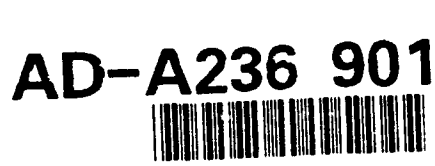


(2)

FINAL TECHNICAL REPORT FOR FEW-ELECTRON LATERAL RESONANT TUNNELING SEMICONDUCTOR DEVICES CONTRACT NO. N00014-89-C-0091

15 April 1991

Prepared for

Office of Naval Research
800 N. Qunicy Street
Arlington, VA 22217-5000

Prepared by

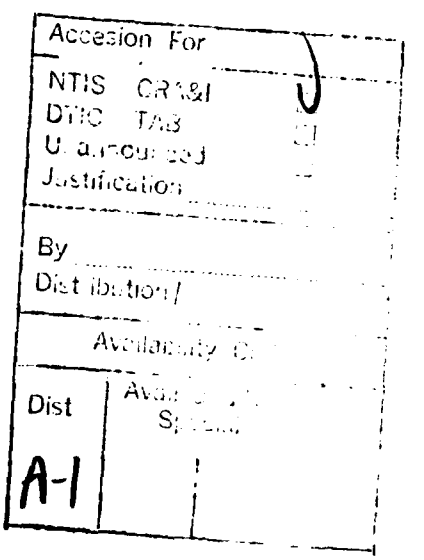
Texas Instruments
13500 N. Central Expressway
Dallas, Texas 75265

FINAL TECHNICAL REPORT FOR FEW-ELECTRON LATERAL RESONANT TUNNELING SEMICONDUCTOR DEVICES CONTRACT NO. N00014-89-C-0091

15 April 1991

Prepared for

Office of Naval Research
800 N. Qunicy Street
Arlington, VA 22217-5000



Prepared by

Texas Instruments
13500 N. Central Expressway
Dallas, Texas 75265

REPORT DOCUMENTATION PAGE	1. REPORT NO.	2.	3. Recipient's Accession No.

4. Title and Subtitle

Few-Electron Lateral Resonant Tunneling Semiconductor Devices

5. Report Date
April 1991

6.

7. Author(s)

R.T. Bate, J. Luscombe, W.R. Frensley, J.N. Randall, M.A. Reed, A. Seabaugh

8. Performing Organization Rept. No.
08-91-09

9. Performing Organization Name and Address

Texas Instruments Incorporated
13500 N. Central Expressway, M/S 105
Dallas, Texas

10. Project/Task/Work Unit No.

11. Contract (C) or Grant (G) No.
(C) N00014-89-C-0091
(G)

12. Sponsoring Organization Name and Address

Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

13. Type of Report and Period Covered
Final Technical Report
June 1989—February 1991

14.

15. Supplementary Notes

16. Abstract (Limit 200 Words)

We report here the results of a contract carried out at Texas Instruments and Yale University. The effort included design, modeling, fabrication, and characterization of lateral resonant tunneling and quantum point contact structures. Also included was a theoretical investigation of open quantum systems driven far from equilibrium, with emphasis on appropriate boundary conditions for solution of such systems. Accomplishments under this contract include the publication in *Reviews of Modern Physics* of the results of this foundational study, as well as the development of a graphics-oriented program for the computation and display of two-dimensional self-consistent energy band diagrams. The first lateral resonant tunneling transistors to exhibit both negative differential conductance and negative transconductance were demonstrated. The eigenstates of finite superlattices driven below the Stark localization threshold were also observed.

17. Document Analysis a. Descriptors

Resonant tunneling, Lateral tunneling, Quantum systems, Superlattices, Gallium arsenide, Quantum wells, Quantum wires, Quantum dots, Nanoelectronics, Mesoscopic systems, Nanostructures, Quantum point contacts

b. Identifiers and Open-Ended Terms

c. COSATI Field/Group

18. Distribution Statement	19. Security Class (This Report)	21. No. of Pages
	Unclassified	20
	20. Security Class (This Page)	22. Price
	Unclassified	

TABLE OF CONTENTS

LIST OF APPENDIXES

FIGURES

FIGURES
(Continued)

TABLES

SECTION I
INTRODUCTION

A. NEED FOR QUANTUM DEVICES

Downscaling of transistor-based IC minimum geometries will eventually be brought to an end by a combination of problems related to devices, interconnections, noise, and reliability.[1] The resulting saturation of circuit densities almost certainly implies a saturation of the historical exponentially downward trend in cost and volume per bit or function, which has been a primary driving force for the increasing pervasiveness of electronics in DoD systems. Scaling has also provided exponential improvements in device speed and power dissipation, which has led to substantial enhancement of system performance. Because the introduction of sophisticated electronics into these systems has significantly improved their capabilities, it is appropriate to determine if there is an alternative scenario that significantly prolongs exponential trends in cost and performance.

Estimates based on abstract physical device switching models that are independent of specific device technologies indicate it would be theoretically possible to achieve several orders of magnitude improvement in downscaling of device powers in devices with minimum geometries of a few hundred angstroms if we could find an appropriate nonconventional transistor device technology. The key to this search is to use electronic phenomena that are characterized by dimensions much smaller than the depletion layer widths and diffusion lengths that provide the basis for conventional transistor function.

A step can be taken in this direction by employing heterojunctions rather than p-n junctions to introduce potential barriers for carrier confinement. The advent of MBE and similar technologies permits us to fabricate semiconductor heterostructures with features on the scale of nanometers. This allows us to explore novel physical phenomena enabled by nanoscale heterostructures that can lead to truly revolutionary device mechanisms. Because semiconductor structures having dimensions comparable to the Bloch wavelength of electrons can be fabricated, the obvious place to look for such phenomena is in quantum-mechanical effects.

B. TUNNELING AS GENERIC QUANTUM EFFECT

The seminal work of Esaki and Tsu,[2] proposing the first artificial semiconductor superlattice, was instrumental in motivating researchers to bandgap engineer semiconductor systems. These authors proposed that such structures would exhibit negative differential conductivity because of the creation of artificial minibands and minigaps. The resonant tunneling diode was realized four years later (1974) by Chang, Esaki, and Tsu.[3] Initial observations were of weak structure in current-voltage characteristics because of resonant tunneling through a single quantum well encased by tunnel barriers; in essence, a single component of the superlattice. More recent work by Sollner et al.[4] revived interest in these devices when peak-to-valley current ratios as high as 6 were observed at 25 K with high-frequency current response (exceeding 2.5 THz).

As an effect, quantum-mechanical tunneling becomes important when the thickness of the potential barrier is on the order of the electron wavelength. This effect provides an alternative means by which charge transport in electron devices can be controlled. In the vast majority of semiconductor switching devices, thermionic or diffusive current transport is controlled by

modulating the potential between device input and output. An alternative method for switching in semiconductor devices is by control of resonant tunneling transmission resonances using tunneling heterostructures.

Research at Texas Instruments (TI) and elsewhere on tunneling devices has been intense over the past years with the most notable achievements being the demonstrations of quantum dots,[5] quantum-well base resonant tunneling transistors,[6,7] and the demonstration of resonant tunneling transistors operating at room temperature with both dc and microwave gain.[8]

C. LATERAL TUNNELING DEVICES

Lateral resonant tunneling devices were first proposed by Bate[9] and Sakaki.[10] Interest in lateral resonant tunneling structures stems from their unique capability for external control of the tunnel barrier height and for their longer elastic and inelastic scattering times compared to vertical resonant tunneling devices, since transport occurs in the plane of a high-mobility two-dimensional electron gas.

Until recently, the ability to create structures having dimensions comparable to the Bloch wavelength of electrons has been restricted to the vertical, epitaxial-growth direction. However, recent advances in nanolithography have allowed researchers to define lateral dimensions comparable to the vertical dimensions mentioned previously. Lateral confinement greatly extends our ability to control electron energy levels in semiconductor devices. TI has been a leader in the development of quantum devices using lateral, as well as vertical, confinement of carriers. The quantum-dot resonant tunneling diode, in which resonant tunneling is controlled by a discrete set of energy levels in a laterally confined quantum well, was first fabricated and electrically characterized at TI.[5] We note that conventional RTDs, without the additional lateral confinement of quantum-dot diodes (QDD), operate by reducing the density of states of a tunneling carrier from three dimensions in the emitter and collector to two dimensions in the quantum well, or 3-2-3. However, the QDD with lateral confinement achieves much sharper NDR characteristics by reducing the density of states of a tunneling carrier to 1-0-1.

During this contract, we extended our ability to use lateral dimensions in the control of carrier transport at nanometer-length scales through development of a lateral resonant tunneling transistor (LRTT). Figure 1 shows such a structure. In this embodiment, electron-beam lithography is used to produce gates on top of a GaAs/AlGaAs modulation-doped two-dimensional electron gas (2DEG). While such structures have been fabricated previously in the AlGaAs/GaAs heterojunction system,[11] in this work we use the improvements obtained in the InAlAs/InGaAs 2DEG. The energy band profile from source to drain forms a double-barrier/single quantum-well structure with lateral quantum-well states formed between the barriers. Applying a monotonically increasing source/drain bias brings the lateral states through energetic alignment with the source states with a resultant negative differential resistance (NDR).

Quantum point contact test structures were also developed during this contract. Quantum point contacts[11] can provide for additional density-of-states reduction in lateral tunneling carriers, which could have novel device applications. Here, one expects, because of two closely separated gates, that conduction in such "contacts" will occur in subbands. As the contact width is electrostatically decreased (increased), a subband channel is removed (added) with a corresponding decrease (increase) in conductance. The key point is that the lateral conductance is quantized.

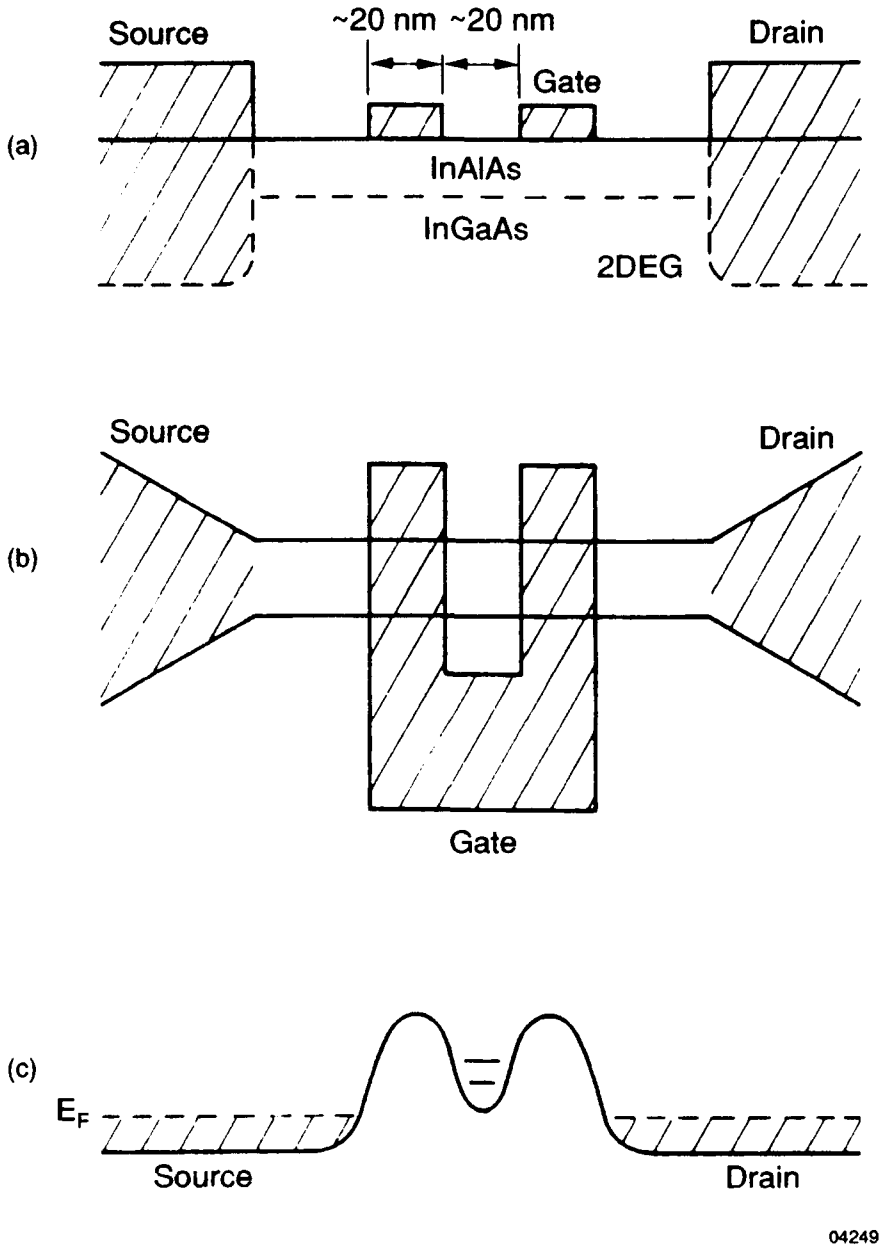


Figure 1
A Lateral Resonant Tunneling Diode. (a) Vertical Physical Structure; E-Beam Written Gates are Formed on Top of a InGaAs/InAlAs Modulation-Doped 2DEG. (b) Top View of The Lateral Resonant Tunneling Transistor. (c) Energy Band Profile From Source to Drain ($V_{sp} = 0$) of the Lateral Resonant Tunneling Potential. The Gate Forms a Double-Barrier/Single Quantum-Well Structure, With Lateral Quantum-Well States Formed Between the Barriers.

In Section II, we describe modeling activities undertaken during this contract and also the relevant design criteria for the LRTT. In Section III, we review the essential elements of the nanofabrication, and in Section IV, we describe the progress on the LRTTs, quantum-point contact structures, and a parallel effort to understand the transition from superlattice miniband to coupled-well structures. We describe the first demonstration of a lateral resonant tunneling transistor, formed using depletion tunnel barriers, which exhibits NDR and negative transconductance. The superlattice miniband devices were realized using vertical tunneling heterostructures, but have application to both vertical and lateral devices. Our summary and conclusions are in Section V.

SECTION II
QUANTUM TRANSPORT THEORY AND DEVICE DESIGN

A. BOUNDARY CONDITIONS FOR OPEN QUANTUM SYSTEMS

The worldwide theoretical effort to describe the behavior of tunneling devices is fragmented into a number of different approaches, each of which focuses on a particular aspect of the problem. The most popular approach invokes the formal theory of scattering to construct wavefunctions that asymptotically approach traveling waves.[13] This is a good way to evaluate steady-state behavior of a system in which transport is purely ballistic, but it is not adapted to address the problems of either transient behavior or dissipative transport. The means by which these issues are addressed, relating the resonance width to the characteristic time or adding a term to the resonance width to approximate inelastic scattering, are clearly inadequate.

An approach that addresses the significant problems of tunneling devices is a quantum kinetic transport theory developed during the course of this contract. In this approach, the mixed state of a quantum system is represented by the Wigner distribution function or an equivalent density matrix. The time evolution of the Wigner function is described by a quantum kinetic equation that incorporates the coupling of the device to its contacts and that can include realistic collision terms. This study culminated in the publication of "Boundary Conditions for Open Quantum Systems Far From Equilibrium," in *Reviews of Modern Physics*[14] and is included as Appendix I.

B. TWO-DIMENSIONAL ENERGY BAND COMPUTATIONS FOR LATERAL HETEROSTRUCTURE DEVICE DESIGN

The second theoretical task of this contract of more direct concern to the experimental program was to provide modeling tools that relate the structural design of the device to its electrical characteristics. At the inception of this work, there were modeling codes in place for the self-consistent band-edge profile for vertical RTDs in which the current flow is parallel to the growth direction of the epitaxial layers. For such devices, a one-dimensional calculation is sufficient, since the device is practically uniform in the remaining two dimensions. The focus of this program, however, was to develop lateral quantum-device technologies in which current flow is perpendicular to the growth direction and parallel to the heterointerfaces that form the 2DEG. Hence, to accommodate the extra, lateral dimension, a two-dimensional simulation capability was required to guide the development of LRTTs.

We, therefore, developed a general two-dimensional device simulation code, NANO2D, that obtains the self-consistent potential energy surface defined by the conduction band minimum for a wide class of two-dimensional III-V semiconductor heterostructure devices. Using a finite-temperature Thomas-Fermi approximation for the carrier density, the solution to a self-consistent two-dimensional nonlinear Poisson equation is obtained for specified contact voltages. This approach is also known as a "zero-current" approximation in device literature. We have demonstrated that the zero-current approximation works well in lateral nanostructure devices.[15-18] The key idea is that the tunnel barriers, in essence, separate a device into regions in which an approximately equilibrium electron distribution is established, and the Thomas-Fermi expression for the carrier density function is the local thermodynamic equilibrium approximation. By having a picture of the two-dimensional potential energy surface, we can accurately predict the quantum energy levels that control current flow. By "two-dimensional" we mean that the user can specify,

in addition to an arbitrary sequence of epitaxial layers in the vertical direction, an arbitrary lateral variation of material composition and doping, such as might be achieved by regrowth or implant techniques. In addition, the user can specify the lateral bias across ohmic source and drain contacts, as well as voltages applied to a back ohmic contact, and to one or more independently contacted top Schottky gates. NANO2D then generates a three-dimensional graphical image of the conduction band minimum. Included as Appendix II is the documentation for using NANO2D.

We illustrate the LRTT device using NANO2D in Figure 2. A particular InGaAs/InAlAs heterostructure, listed in Table 1 and described further in Section IV, is chosen to illustrate the device model. Considering Figure 2, the surface of the device is at z = 100 nm and the 2DEG at the InAlAs/InGaAs interface is at z = 58 nm. The device is contacted as in a conventional modulation-doped field-effect transistor; however, in this device the potential of the 2DEG channel is modulated by two closely spaced and narrow gates. This modulation is apparent in Figure 2 looking along the z = 58-nm plane. In this device, the gates are approximately 60-nm wide and

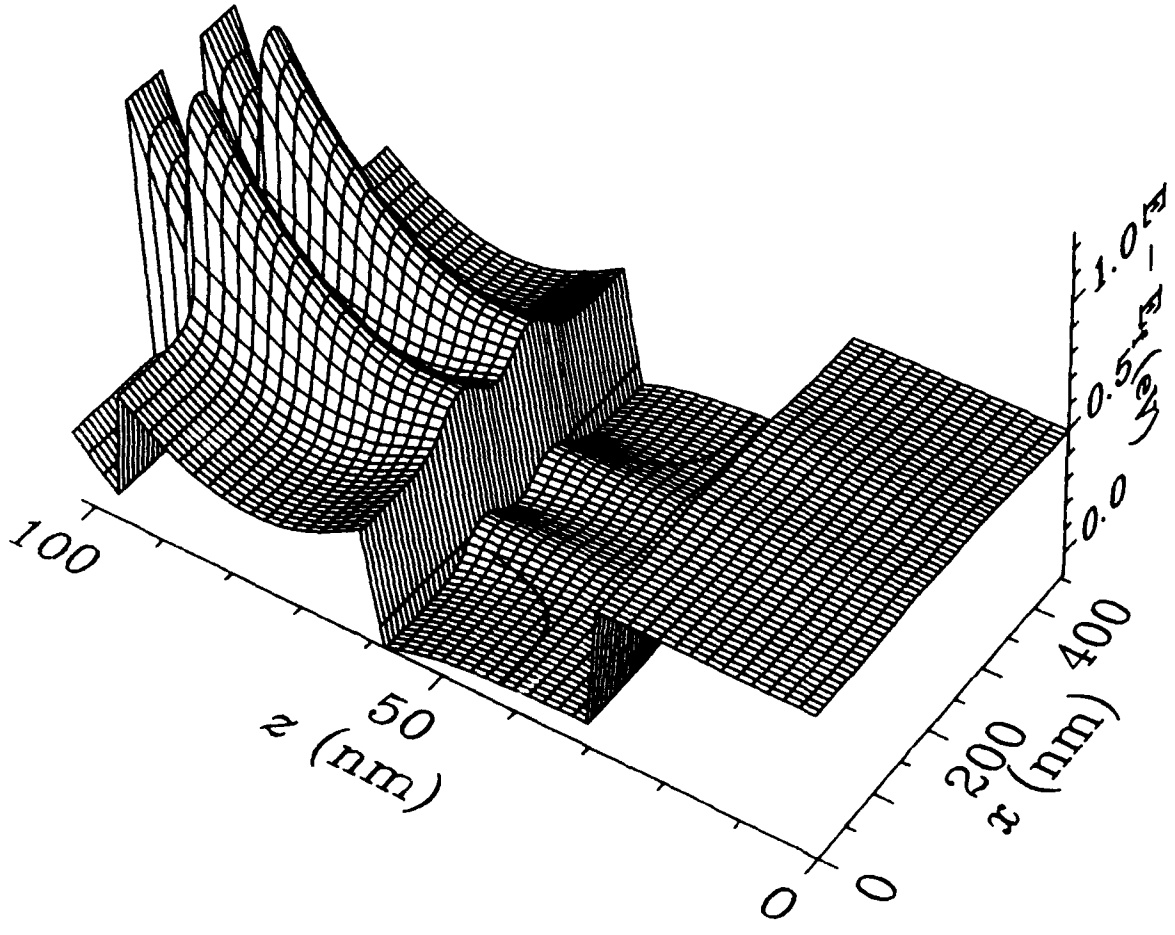


Figure 2
Two-Dimensional Energy Band Diagram for an $In_{0.52}Ga_{0.47}As/In_{0.53}Ga_{0.47}As$ Lateral Resonant Tunneling Transistor Under Zero Lateral Bias, Zero Substrate Bias, and 1-V Applied to Gate Electrodes. Parameters are Listed in Table 1.

separated by 60 nm. These dimensions are of the same order as the Fermi wavelength and introduce size quantization effects into the device characteristics, as will be shown in Section IV. The lower InAlAs buffer layer is inserted to reduce parasitic parallel conduction, which can be significant in InGaAs.

Table 1. Material and Device Structure Used in the Numerical Simulations of Wafer R5023.

Thickness (nm)	Composition	Doping (cm^{-3})
8	$In_{0.53}Ga_{0.47}As$	1×10^{18}
30	$In_{0.52}Al_{0.48}As$	1×10^{18}
5	$In_{0.52}Al_{0.48}As$	—
30	$In_{0.53}Ga_{0.47}As$	—
600	$In_{0.52}Al_{0.48}As$	—
Substrate	InP	semi-insulating
Lateral gate length	60 nm	
Gate spacing	60 nm	
Fermi-level pinning	0.2 eV	
Gate surface potential	1.0 eV	

A more conventional one-dimensional view of the epitaxial structure, from the center of the gates extending in the z-direction from the surface toward the substrate, is shown in Figure 3. Clearly, for this material design, surface layers are depleted and the only conduction path below the Fermi-level is along the 2DEG. Finally, the energy band profile for the device in the plane of the 2DEG shows the two depletion barriers induced by the surface gate contacts (Figure 4). Note that the potential well formed between the barriers is approximately the harmonic oscillator potential, for energies less than the barrier height.

With source and drain connections to either end of the 2DEG channel, electrons traveling from source to drain encounter two depletion tunnel barriers analogous to the heterojunction double barriers obtained in vertical RTDs. Unlike vertical devices, tunneling occurs from high-mobility 2-D electrons, through 1-D confined well states, to 2-D drain states. Also, unlike the vertical RTD, tunnel barrier heights are field-controllable, introducing an additional degree of freedom to current-voltage spectroscopy.

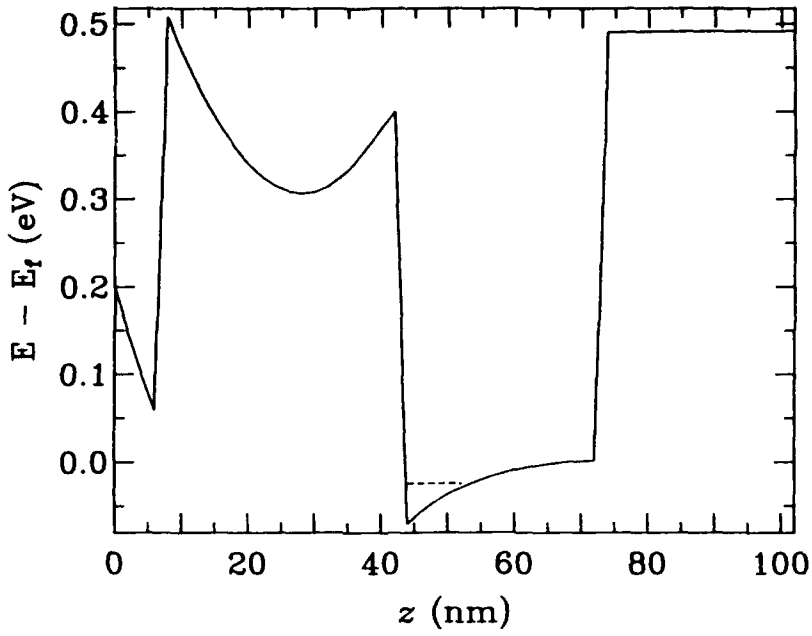


Figure 3
Computed Energy Band Profile in the Growth Direction for the InAlAs/InGaAs 2DEG Device of Figure 2

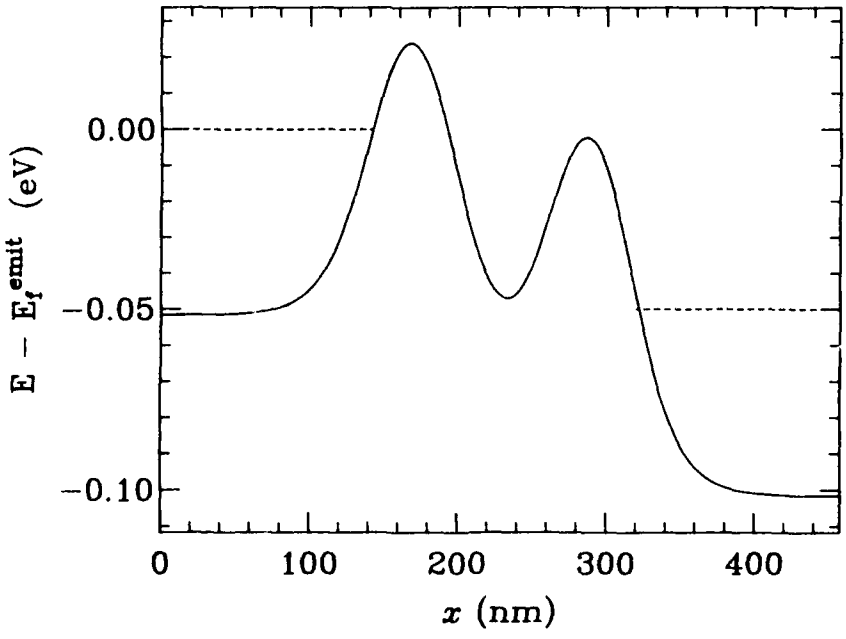


Figure 4
Computed Energy Band Profile Along the 2DEG at a Position 5 nm From the InAlAs/InGaAs Near the Wave-Function Peak for the Same Device Described in Figure 2

SECTION III
NANOFABRICATION

A. LAYOUT

We fabricated all three devices, with the same general process, on a single die that was 1.4 mm by 1.1 mm. A mask set, shown in Figure 5, was designed that contained the following five levels: alignment marks, mesa, ohmic metal, gate metal, and pads. The alignment level was defined in metal and contained both optical and e-beam alignment marks. The devices were designed so that the alignment accuracy required for the optical lithography steps was very easily met while we took advantage of the high resolution and excellent overlay accuracy of our Philips e-beam lithography tool.

The mesa level is an e-beam exposed level. Although the required resolution of 0.7 μm is not out of the question for optical lithography, the alignment accuracy called for is. The ohmic metal level is exposed with optical lithography.

The gate level is formed with a combination of e-beam and optical lithography. The minimum feature for the QPC and spectrometer devices is 0.1 μm, while the lateral tunneling device requires sub-0.1-μm lines. The optical portion of the gate metal mask shorts all of the bondpads to separate gate electrodes together. This is to avoid damage caused by even small electrostatic charges that the device might be subjected to. Once the device was packaged and bonded, the metal lines shorting the pads together are scribed and shorts are removed. There is also an optional optical mask pattern for vias through a dielectric passivating layer and/or adding additional metal to the bondpad area.

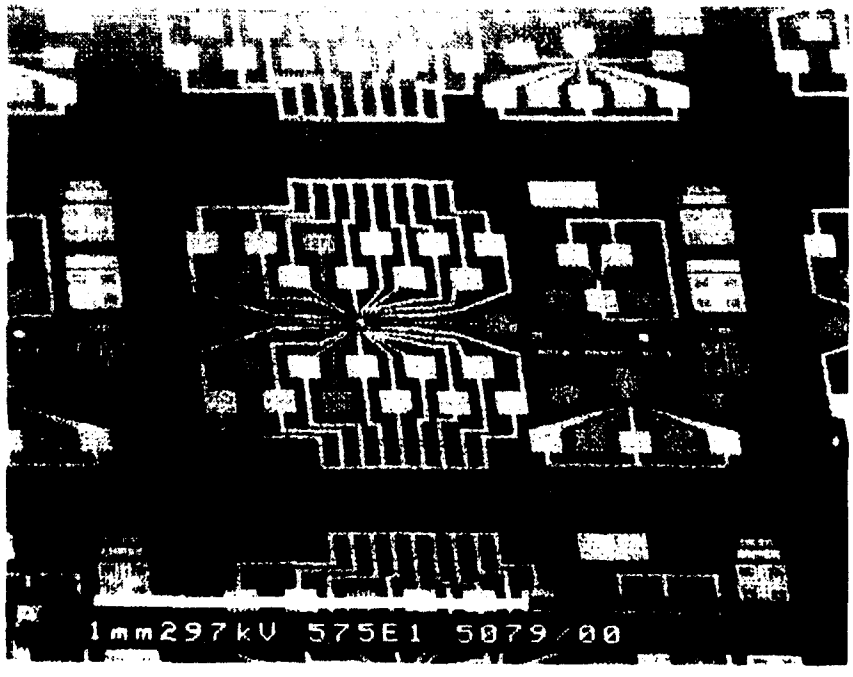


Figure 5
Scanning Electron Micrograph of Completed Devices
Produced by the Mask Set Developed for This Program

B. PROCESS

Postepitaxial fabrication begins with definition of the alignment mark pattern by optical lithography, where pattern transfer by vacuum evaporation and lift-off yields a metal pattern. Metallization was typically Cr or Ti, followed by approximately 200 nm of Au.

Either the mesa or ohmic levels could be formed next. There are advantages to either approach. Patterning the ohmic metal first allows the use of an electrical measurement to determine the effectiveness of the mesa etch in isolating the 2DEG. This ability is desirable when one wishes to use the minimum depth etch for isolation so that step coverage problems are reduced. However, metal etch masks sometimes lead to anomalously high etch rates adjacent to the metal. This can result in undercutting the mesa area, disconnecting it from its ohmic contacts. Although both processes were tried, we more typically patterned and etched the mesa areas before ohmic metallization.

The mesa pattern called for a negative e-beam resist. We experimented with both Shipley SAL-601 and CMS-EXR. For our particular processing needs, we found the CMS-EXR to be superior. Using a dose of 70-$\mu C/cm^2$ with a 50-keV beam produced excellent pattern fidelity on both GaAs and InP substrates. After a postdevelopment bake, samples (both GaAs and InGaAs) were etched with sulfuric acid, hydrogen peroxide, and water mixtures in the ratio 1:8:160. In our epitaxial structures, the 2DEG was relatively close to the surface so that we could etch to a depth that ensured isolation without experiencing step coverage problems.

Negative e-beam resist is difficult to remove by any other method than O_2 plasmas. We did have some concern that the relatively long plasma etching times required to remove this resist might have a deleterious effect on 2DEG mobility. To determine this, van der Pauw/Hall-effect samples were subjected to O_2 plasma etch conditions for 30 minutes. We detected no significant degradation in mobility of these samples after plasma ashing. For GaAs samples, ohmic metallization consisted of AuGe/Ni/Au metallurgy, followed by a furnace anneal at 430°C for 3 minutes. InGaAs samples received Cr/Au or Ti/Pt/Au ohmic metallizations and were alloyed similarly. In both cases, a simple optical lithography step, followed by lift-off, accomplished the pattern transfer.

To reproduce the designed gate pattern faithfully, a fairly wide parameter space of exposure conditions had to be explored on our e-beam. The QPC and spectrometer gate patterns were produced with a 50-nm-diameter beam, 25-nm pixels, and a dose of 450 $\mu C/cm^2$. Polymethylmethacrylate (PMMA) with a molecular weight of 950,000 was used as a resist. After exposure, samples were spray developed with MIBK/Ipa 1:1 for 2 minutes. Vacuum evaporation and lift-off procedures were used to form the metal gate pattern. Figures 6(a) and (b) and 7(a) and (b) show the mesa and gate pattern design and SEMs of fabricated QPC and spectrometer devices, respectively.

To form electrostatic tunnel barriers for the lateral resonant tunneling transistor, we employed a 15-nm-diameter beam with 5-nm pixel spacing. Because of intraproximity effects, the required dose is much higher: 2600 $\mu C/cm^2$. Figure 8 is an SEM of the dual-gate structure produced. Each gate is approximately 60-nm long with a 60-nm space between them.

A passivating layer of silicon nitride was plasma deposited on the GaAs samples and vias were etched through to the bondpads. On InGaAs samples, we noticed a reduction in sample conductivity. We suspect that the nitride layer may affect the surface potential of InGaAs. In subsequent samples, we did not deposit silicon nitride on InGaAs devices.

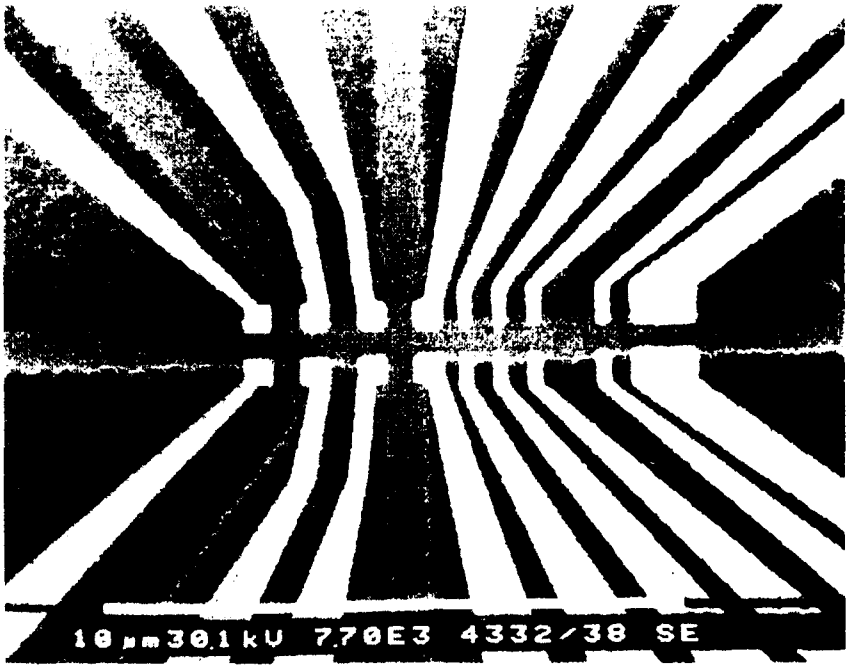


Figure 6
(a) Close Up View of the Design of Mesa and Gate Levels for Quantum Point Contact Structure. (b) Scanning Electron Micrograph of Fabricated Quantum Point Contact Structure.

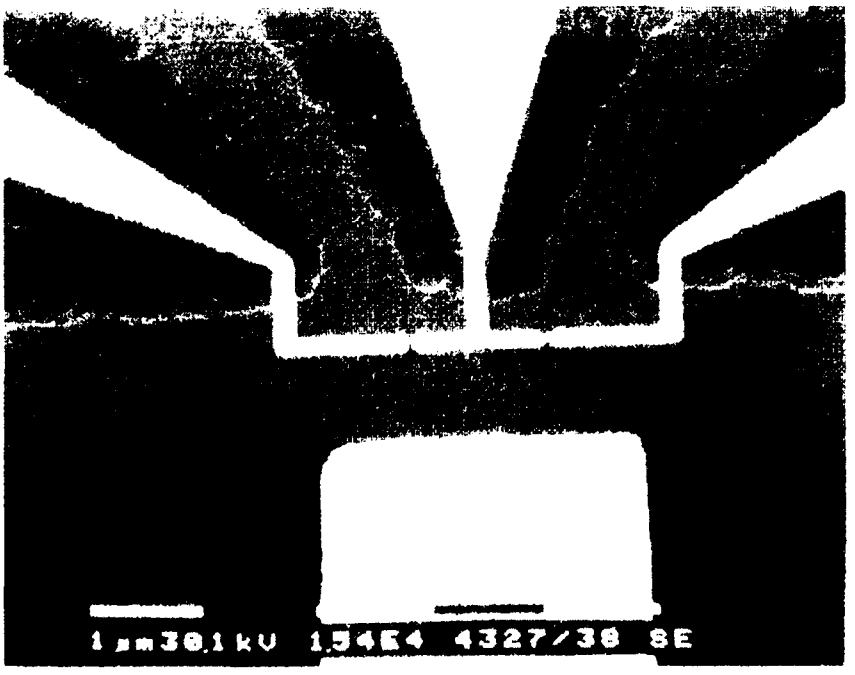


Figure 7
(a) Close Up View of the Design of Mesa and Gate Levels for Electron Spectrometer Structure. (b) Scanning Electron Micrograph of Fabricated Electron Spectrometer Structure.

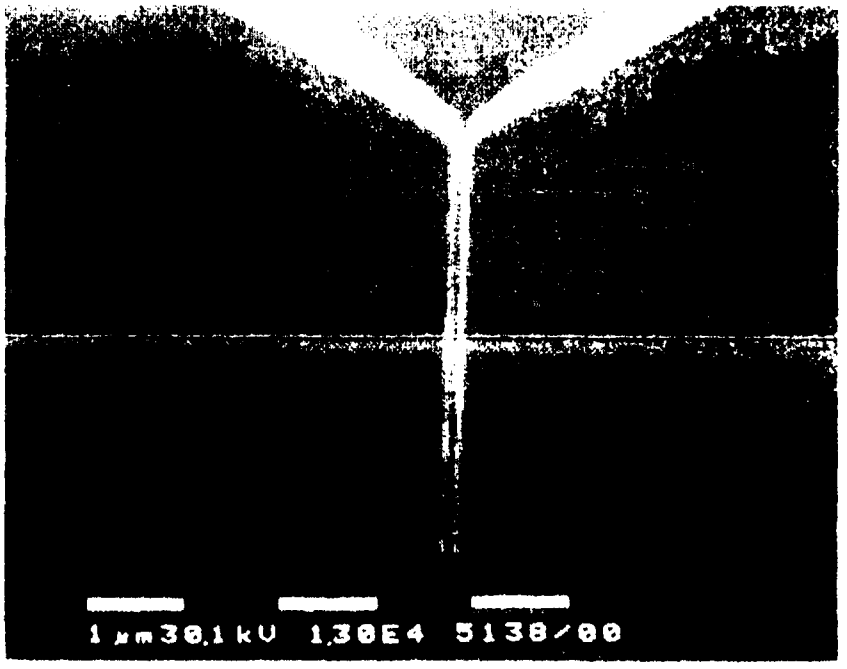


Figure 8

Scanning Electron Micrograph of Two-material Coated
Lateral Resonant Tunneling Transistor

SECTION IV
CHARACTERIZATION & ANALYSIS

A. LATERAL RESONANT TUNNELING TRANSISTOR

A lateral resonant tunneling field-effect transistor, similar to that described in Section II, has been previously demonstrated in the AlGaAs/GaAs system by Ismail et al.[11] Ismail showed clear conductance oscillations associated with resonant tunneling of electrons across the depletion barriers at 4.2 K; however, no NDR was observed. In this section, we describe the first observation of NDR in a lateral resonant tunneling transistor. The NDR in this InAlAs/InGaAs lateral RTD persists to 30 K. Furthermore, we show clear evidence for mode mixing of 2-D and 1-D confined states.

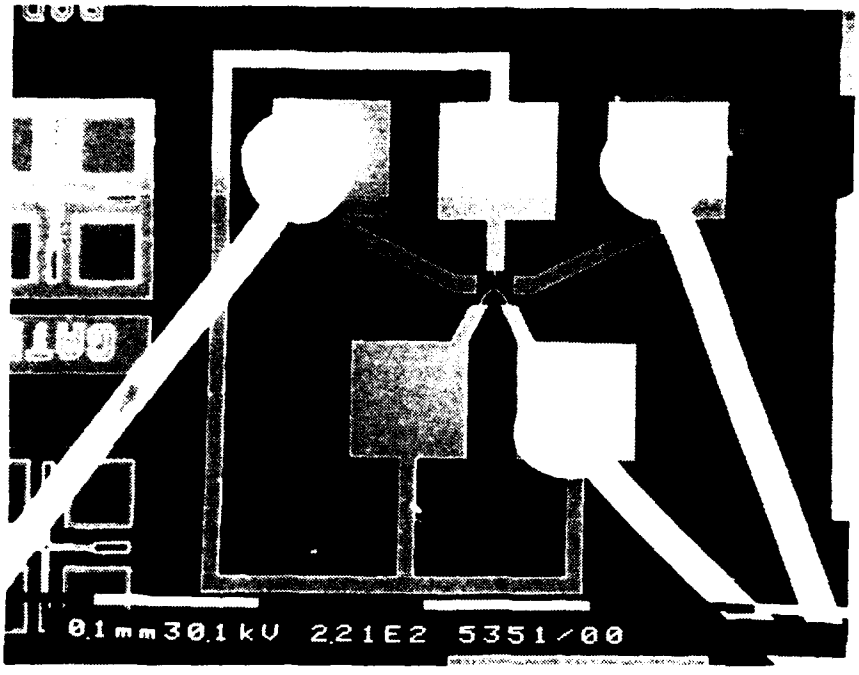
The material structure for this device was described in Section II. Van der Pauw resistivity and Hall-effect measurements for the heterostructure, R5023, yield a room-temperature sheet carrier density of 1.12×10^{12} cm^{-2} with a mobility of 9500 cm^2/Vs. At 77 K, sheet carrier density is 9.6×10^{11} cm^{-2} with a mobility of 49800 cm^2/Vs. The device structure consists of an 0.7-μm-wide mesa with source-to-drain spacing of 20 μm. Dual 60-nm gates, spaced 60-nm apart overlie the mesa, as shown in Figure 9. For the device described, gates are not connected externally. Modulation of the 2DEG is obtained by bias applied to the substrate backside, V_S.

These devices exhibit persistent photoconductivity (PPC) when illuminated briefly by a red light-emitting diode. The PPC is characterized by two time constants at 4.2 K: an initial short decay constant of about 2 minutes, and a longer decay time exceeding several hours. Measurements described here were made after the short-lived photoconductive decay. No significant conduction is obtained without prior illumination. Presumably, the PPC effect is similar to that observed in other InP/InGaAs and InAlAs/InGaAs 2DEG heterostructures.[19,20] This PPC occurs because of separation of photogenerated electron-hole pairs or donor photo-ionization and charge separation at the 2DEG heterojunction.

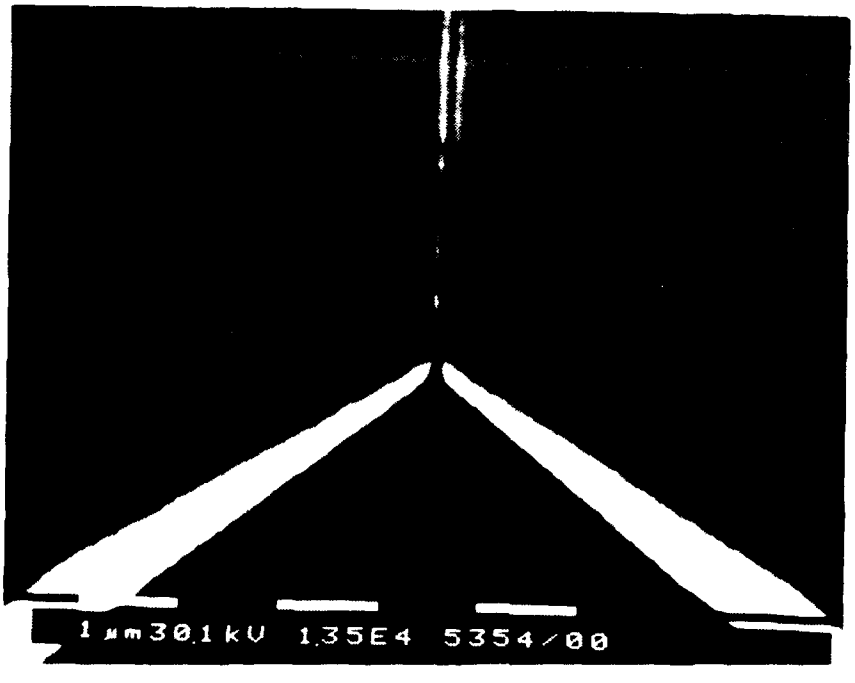
Detailed I-V characteristics of this lateral resonant tunneling device are shown in Figure 10 as a function of substrate bias. As substrate bias is increased, the Fermi-level is raised with respect to the depletion barriers. Because the barrier shape is quasi-Gaussian, the highest transmission probability occurs at the Fermi-energy, where effective barrier thicknesses are thinnest. We observe resonant tunneling as a function of drain-source bias with as many as seven peaks discernable (e.g., $V_S = 2.5$ V, reproduced alone in Figure 11).

Note that the peak separations are nearly equal, as would be expected from a symmetric harmonic oscillator potential. Using the one-dimensional computed potential profile as in Figure 4, an estimate for expected peak-voltage separations can be obtained from symmetric harmonic oscillator eigenvalues. For this potential, $\hbar\omega$ is approximately 14 meV, in good agreement with the value of 17 meV observed in these measurements.

Also in Figure 10, the first two resonant tunneling peaks are observed to successively disappear as substrate bias is increased. The effect of the substrate bias is to raise the Fermi-energy with respect to the potential barriers. When the quantum-well states are moved below the quasi-Fermi level, transmission through these states is apparently significantly suppressed because of their occupancy and coulomb repulsion. The general shift in peak positions with bias is likely caused by the increase in channel conductance with substrate bias.



(a)



(b)

Figure 9
SEM Micrographs of the InGaAs/InAlAs Lateral RTD: (a) Overall View of Bonded Device With Two Upper Bonds Connected to Source and Drain and Lower Bond Connected to Gate. (b) Expanded View of Dual 60-nm Gates Crossing 0.7-μm-Wide Mesa

Another notable feature in Figure 10 is a slow charging effect in the device with a time constant less than one minute. In the measurement of Figure 10, the drain source voltage is stepped from 0 to 0.2 V at one substrate bias, followed by a second sweep of V_{ds}, and so on, to complete a set of five substrate biases (limitation on data set size is given by the HP semiconductor parameter analyzer used to make these measurements). It is apparent in Figure 10 that, for substrate biases exceeding 2.7 V, the first measurement in these data sets (indicated by the dashed lines) is shifted with respect to the rest of the measurements in the set. This is indicative of a charging effect associated with the applied biases. Since the effect is not observed until substrate bias exceeds 2.7 V, it is likely to be associated with trapping in the InAlAs buffer layer or buffer-layer interfaces. We note that the substrate current is small, even at high applied bias, i.e., less than 200 pA at $V_s = 5$ V.

Temperature dependence of resonant tunneling is shown in Figure 12. As can be seen, NDR persists to temperatures as high as 30 K. Evaluation of the I-V characteristic between 1.2 and 4.2 K, and the reason for loss of higher voltage peaks, is not presently understood.

In addition to these measurements at fixed substrate bias, one can also use substrate bias to modulate the channel at fixed drain/source bias. This is fundamentally different from previously described measurements, since in the former case (with fixed substrate bias) electrons are injected from the same subband in the source as the drain/source bias. Under this condition, the spectroscopy is then indicative of the quantum-well eigenvalues that, in our experimental data, have approximately equal separation (no strong effect of well-to-drain selection is indicated). However, under fixed V_{ds}, the 2-D subband occupation in the drain and source are controllable by the substrate bias. Thus, the 2-D source/drain size quantization is revealed in a rich spectrum of resonances between source and drain (see Figure 13). Further characterization and analysis of these results are in progress and will be reported elsewhere.[21]

B. FINITE SUPERLATTICES

Recognizing that issues of electron coherence in superlattices are important in understanding lateral resonant tunneling-device functions, as a secondary task of this contract we made a study of electron transport in vertical, epitaxial superlattices, which could then be applied to lateral structures. This work was published[22] and is included as Appendix III.

C. QUANTUM POINT CONTACTS

Measurements on fabricated quantum point contact structures have not, to date, yielded fully functional devices. The initial three sets of devices received at Yale had problems related to gate leakage, submicrometer ohmic contact formation, and surface passivation (InGaAs devices only). Major processing impediments appear now to be solved, as evidenced by the demonstration of lateral resonant tunneling as described elsewhere in this report. A fourth set of 24 devices was received at Yale as this report was being completed, and is in the process of being measured at this writing.

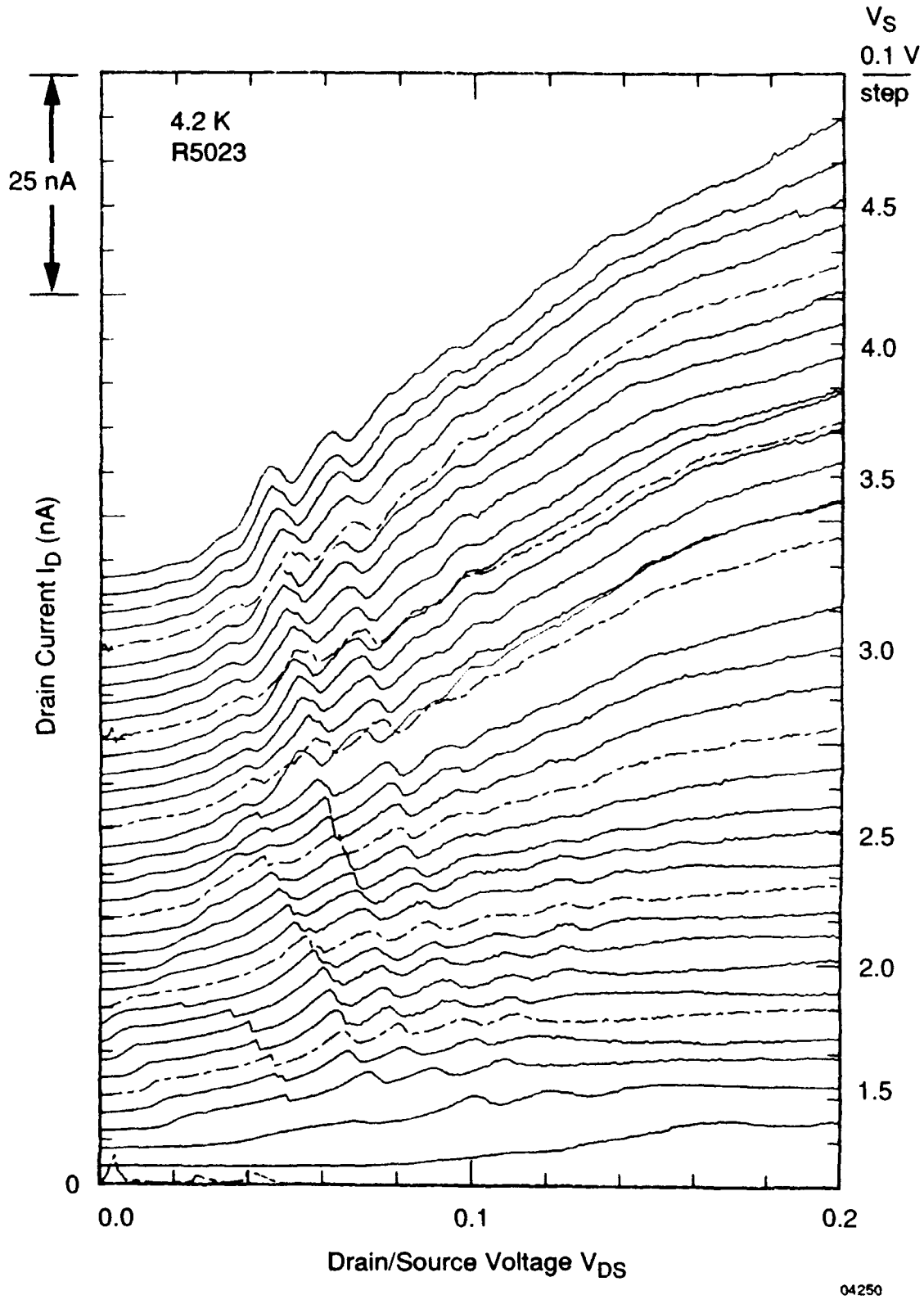


Figure 10
Dependence of Current-Voltage Characteristic on Substrate Bias for an InGaAs/InAlAs Lateral RTD (R5023). For Clarity, Each Trace is Displaced by 2 nA From the Previous Trace. Dashed Lines Indicate the First Trace in a Sequence of Five Separate Substrate Bias.

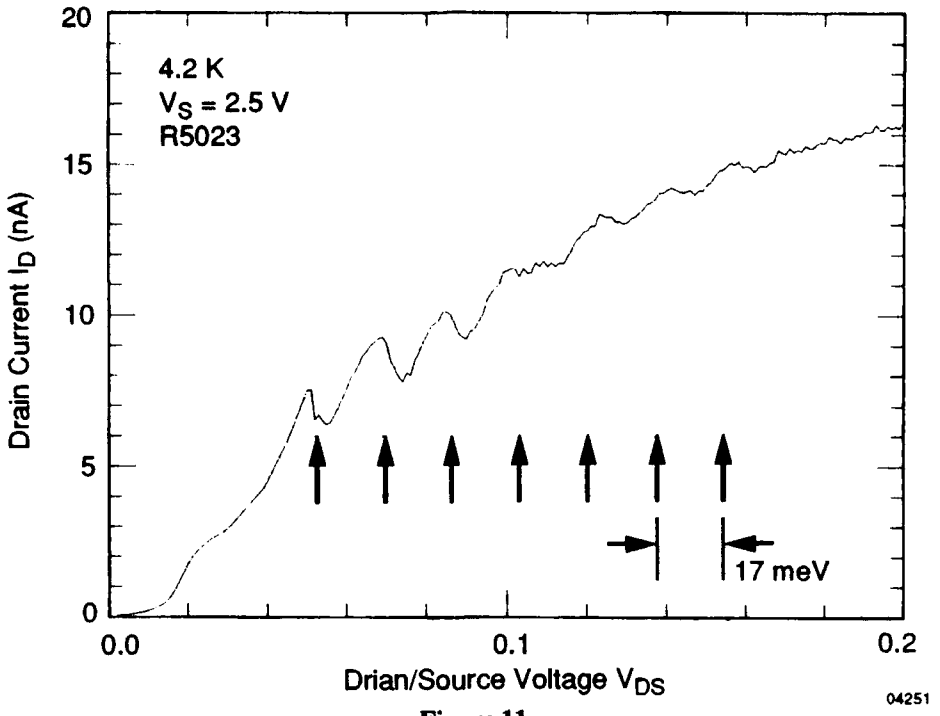


Figure 11
Current Voltage Characteristics at Fixed Substrate Bias, V_s = 2.5 V. Arrows Indicate Expected Peak Voltage Positions for a Symmetric Harmonic Oscillator Potential With Eigenvalue Spacing of 17 meV.

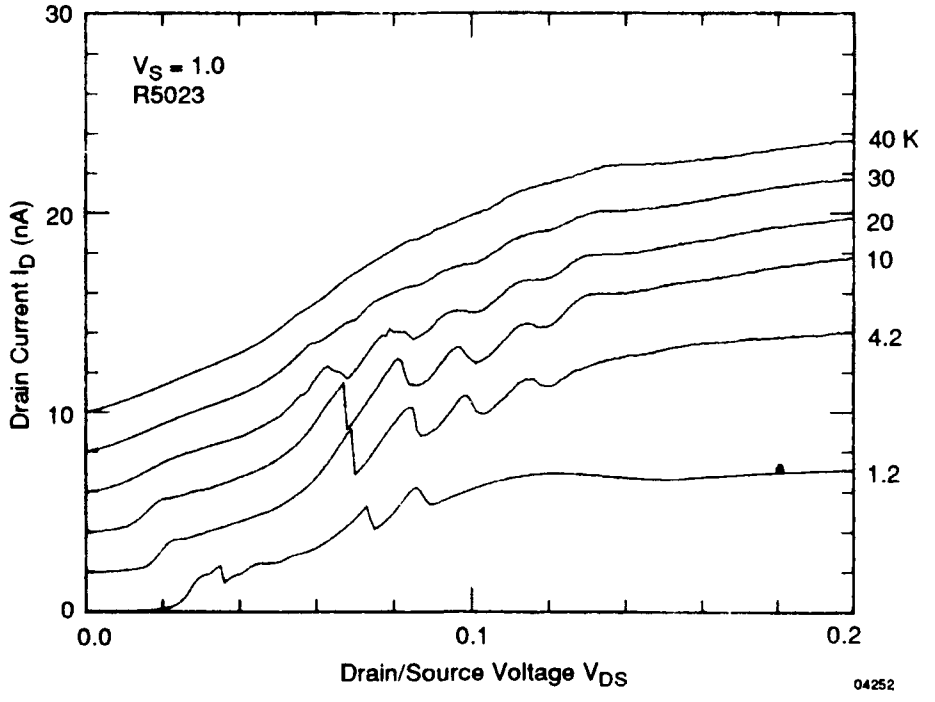


Figure 12
Temperature Dependence of the I-V Characteristics of an InGaAs/InAlAs Lateral Resonant Tunneling Diode for V_s = 1 V (R5023)

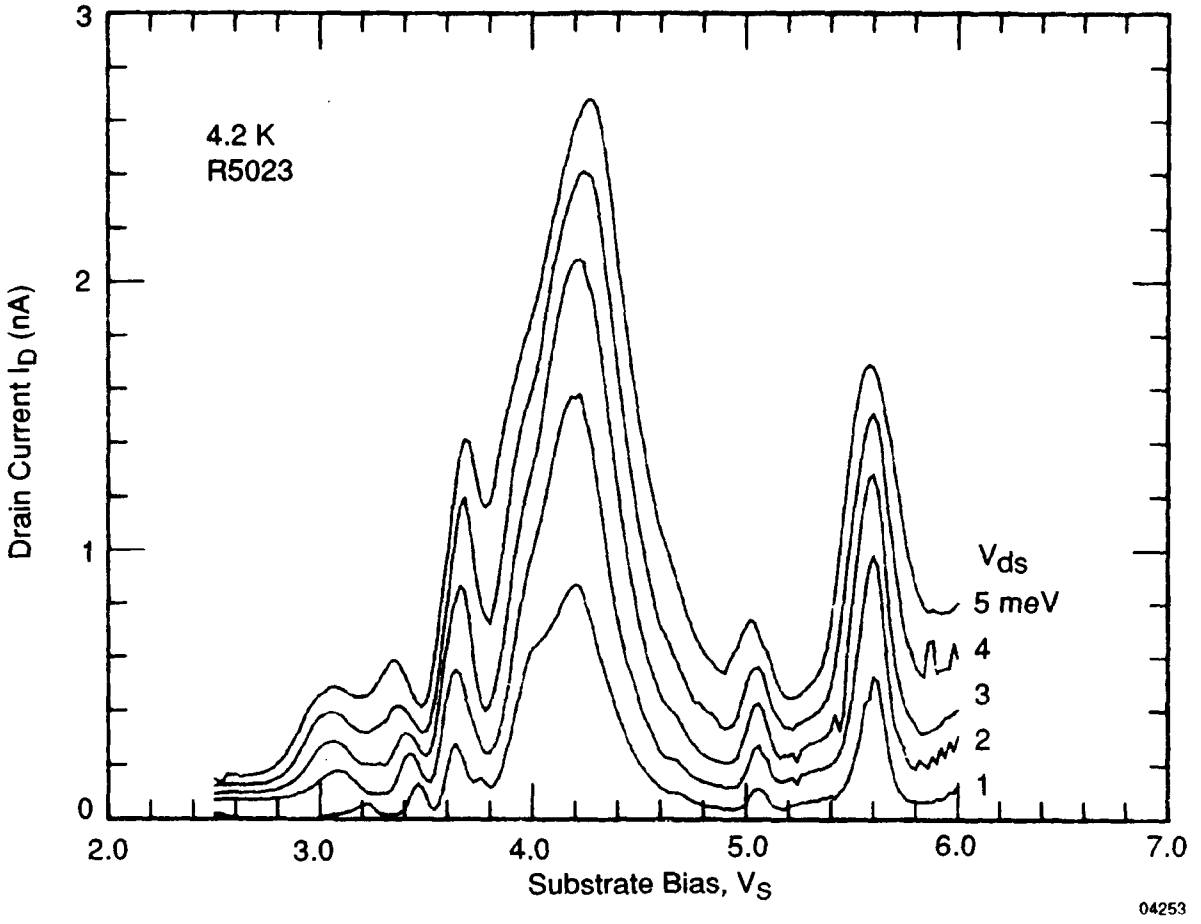


Figure 13
Drain Current Dependence on Substrate Bias for Fixed Source/Drain Voltage at 4.2 K (R5023)

SECTION V
SUMMARY AND CONCLUSIONS

Accomplishments under this contract cover a wide range of activities in an important area of electron device research. We have developed and brought to publication a foundation work[14] (see Appendix I) on the physics of open quantum systems driven far from equilibrium. A parallel modeling effort has enabled the computation of two-dimensional self-consistent energy band diagrams. This code, NANO2D, described in Appendix II, has enabled the successful design and demonstration of the first lateral resonant tunneling transistors to exhibit both negative differential conductance and negative transconductance. Negative resistance is demonstrated to persist to 30 K. Previous work in this field has shown conductance modulation at 4.2 K, but no NDR.

We have also observed, by tunneling spectroscopy of finite superlattices, eigenstates of a superlattice system driven below the Stark localization threshold[22] (see Appendix III).

Significant improvements obtained here call for continued work in this field to examine the feasibility of room-temperature operation of the lateral resonant tunneling device. It is anticipated that such operation is feasible if depletion barriers are replaced by heterojunctions using, for example, an etch and epitaxial regrowth process.

Directions for further development of the modeling area include the ability to compute relevant quantized energy levels to make quantitative predictions for the I(V) characteristics. Since lateral devices lack any simplifying symmetry (such as occurs in cylindrical quantum dots, for example), a multi-dimensional Schrödinger solver would be required in which both the bound-state energy levels of the 2DEG subbands and the resonant tunneling levels are computed in the same calculation. This would permit further understanding of the possible mode mixing effects that occur in the tunneling transitions from a two- to a one- to a two-dimensional electron gas.

REFERENCES

1. R.T. Bate, *Nanotechnology* **1**, 1 (1990).
2. L. Esaki and R. Tsu, *IBM J. Res. Dev.*, 61 (1970).
3. L.L. Chang, L. Esaki, and R.Tsu, *Appl. Phys. Lett*,. **24**, 593 (1974).
4. T.C.L.G. Sollner et al., *Appl. Phys. Lett.*, **43**, 588 (1983).
5. M.A. Reed, J.N. Randall, R.J. Aggarwal, R.J. Matyi, T.M. Moore, and A.E. Wetsel, *Phys. Rev. Lett.*, **60**, 535 (1988).
6. M.A. Reed, W.R. Frensley, R.J. Matyi, J.R. Randall, and A.C. Seabaugh, *Appl. Phys. Lett.*, **54**, 1034, (1989).
7. A.C. Seabaugh, W.R. Frensley, J.N. Randall, M.A. Reed, D.L. Farrington, and R.J. Matyi, *IEEE Trans. Electron. Dev.*, **36**, 2328 (1989).
8. A.C. Seabaugh, Y.C. Kao, J.N. Randall, W.R. Frensley, and A. Khatibzadeh, *Jpn. J. Appl. Phys.*, accepted for publication (1991).
9. R.T. Bate, "Electrically Controllable Superlattice," *Bull. Am. Phys. Soc.*, **22**, 407 (1977).
10. H. Sakaki, K. Wagatsuma, J. Hamasaki, and S. Saito, *Thin Solid Films*, **36**, 497 (1976).
11. K. Ismail, D.A. Antoniadis, and H.I. Smith, *Appl. Phys. Lett.*, **55**, 589 (1989).
12. B.J. van Wees, H. van Houten, C.W.J. Beenakker, J.G. Williamson, L.P. Kouwenhoven, D. van der Marel, and C.T. Foxon, *Phys. Rev. Lett.*, **60**, 848 (1988).
13. B. Ricco and M. Ya. Azbel, *Phys. Rev.*, **B29**, 1970 (1984).
14. W.R. Frensley, *Rev. Mod. Phys.*, **62**, 745 (1990).
15. J.H. Luscombe and M. Luban, *Appl. Phys. Lett.*, **57**, 61 (1990).
16. J.H. Luscombe and W.R. Frensley, *Nanotechnology* **1**, 131 (1990).
17. J.H. Luscombe, J.N. Randall, and A.M. Bouchard, accepted for publication, *Proc. IEEE, Special Issue on Nanoelectronics*, May, 1991.
18. J.H. Luscombe, A.M. Bouchard, and M. Luban, to be submitted, *Phys. Rev. B*.
19. K. Tsubaki, T. Fukui, and H. Saito, *J. Appl. Phys.*, **60**, 3224 (1986).
20. J.M. Kuo, B. Lalevic, and T.Y. Chang, *J. Vac. Sci. Tech.*, **B5**, 782 (1987).
21. A.M. Bouchard, J.H. Luscombe, J.N. Randall, and A.C. Seabaugh, invited talk, International Symposium Nanostructures and Mesoscopic Systems, Santa Fe, NM (1991).
22. R.J. Aggarwal, M.A. Reed, W.R. Frensley, Y.C. Kao, and J.H. Luscombe, *Appl. Phys. Lett.*, **57**, 707 (1990).

APPENDIX I
BOUNDARY CONDITIONS FOR OPEN QUANTUM SYSTEMS DRIVEN FAR FROM EQUILIBRIUM

Boundary conditions for open quantum systems driven far from equilibrium

William R. Frensley*

Central Research Laboratories, Texas Instruments Incorporated, Dallas, Texas 75265

This is a study of simple kinetic models of open systems, in the sense of systems that can exchange conserved particles with their environment. The system is assumed to be one dimensional and situated between two particle reservoirs. Such a system is readily driven far from equilibrium if the chemical potentials of the reservoirs differ appreciably. The openness of the system modifies the spatial boundary conditions on the single-particle Liouville–von Neumann equation, leading to a non-Hermitian Liouville operator. If the open-system boundary conditions are time reversible, exponentially growing (unphysical) solutions are introduced into the time dependence of the density matrix. This problem is avoided by applying time-irreversible boundary conditions to the Wigner distribution function. These boundary conditions model the external environment as ideal particle reservoirs with properties analogous to those of a blackbody. This time-irreversible model may be numerically evaluated in a discrete approximation and has been applied to the study of a resonant-tunneling semiconductor diode. The physical and mathematical properties of the irreversible kinetic model, in both its discrete and its continuum formulations, are examined in detail. The model demonstrates the distinction in kinetic theory between commutator superoperators, which may become non-Hermitian to describe irreversible behavior, and anticommutator superoperators, which remain Hermitian and are used to evaluate physical observables.

CONTENTS

I. INTRODUCTION

The more active, and thus the more interesting, products of technology are systems that operate far from thermal equilibrium. An examination of a few examples of such systems shows that they are generally open, in the sense that they exchange matter with their environment. The present work examines some schemes by which open quantum systems (which are beginning to become technologically important in the context of microelectronics) may be effectively described at a kinetic level.

In the context of the present work, an "open system" is one that can exchange locally conserved particles with its environment. Moreover, we wish to focus upon the far-from-equilibrium behavior of such a system, and thus the definition of *open system* will be further restricted to mean one that is coupled to at least two separate particle reservoirs, so that a nonequilibrium state may be created and maintained. To specify such a system we must regard it as occupying a finite region of space, and thus the exchange of particles must consist of a current flowing through that surface which is taken to be the boundary of the system. It does not appear that the statistical

*Present address: Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, Texas 75083.

physics of such a situation has been the subject of a close examination. [The traditional use of the grand canonical ensemble to define the equilibrium state (Tolman, 1938, Sec. 140) contemplates a system coupled to a single particle reservoir.] There is a large body of work on quantum systems that are coupled to a reservoir so as to permit an exchange of energy (see, for example, Chester, 1963; Louisell, 1973; Haken, 1975; Davies, 1976; Oppenheim, Shuler, and Weiss, 1977; and references therein), or are in purely thermal contact with two or more reservoirs (Lebowitz, 1959). Most of these analyses are directed more to the problem of damping (as seen in ohmic conduction) than to openness in the present sense. Much of the work in this area has been motivated by the development of optical technology (Louisell, 1973; Haken, 1975), in which the present distinction between openness and damping is unnecessary because the particles of interest are massless bosons. In a laser, for example, the degrees of freedom of greatest interest are the normal modes of the radiation field. A single theoretical model, the damped harmonic oscillator, is used to describe both the loss of energy (photons) to the gain medium within the cavity and the loss of photons to the output beam (Gordon, 1967; Scully and Lamb, 1967). The analogous processes in an electronic resistor (an open system in the present sense) are the scattering of an electron by a phonon within the resistive material and the escape of an electron from the resistive material into a more highly conductive contact. The present work will concentrate upon the consequences of the latter process. The difference between the system of massive fermions and the system of massless bosons is that the fermion system is constrained by a local continuity equation, whereas the boson system (within the usual models) is not so constrained.

A. Significance of open systems

To document the importance of open systems, let us consider some examples. Most practical engines (in the sense of machines that convert some form of energy into mechanical work) exchange matter with two or more reservoirs. To cite examples from an earlier technology (avoiding the complications of internal phase transitions or chemical reactions) we might consider the overshot water wheel (Reynolds, 1983), which operates between reservoirs of water at different gravitational potential, or the high-pressure steam engine (Dickinson, 1938), which operates between its boiler and the atmosphere, reservoirs which differ greatly in their pressure and temperature. Conspicuously absent from a list of economically significant engines are systems that operate upon the Carnot model of a closed system in purely thermal contact with its reservoirs.

A technology of more current interest is electronics, whose systems are usually arranged so that a "power supply" maintains constant voltages (i.e., chemical potentials for electrons) on two or more "buses" (see, for example, Horowitz and Hill, 1980). The "circuits" (such as logical gates or analog amplifiers) that perform the intended functions of the system are connected to, and conduct current between, the buses. Each bus is an electron reservoir, and the performance of the system's power supply is judged by how nearly these reservoirs approach the ideal behavior of no change in chemical potential (voltage) as particles are exchanged (current is drawn).

The example of electronics points out that the distinction between a closed and an open system depends upon how one chooses to partition the universe into the system of interest and "everything else." (Such a partitioning is implicit in the analysis of every physical problem.) To demonstrate this point, let us examine the etymology of the term *circuit*. As used in the preceding paragraph, *circuit* means "an assemblage of electronic elements" (Woolf, 1981), which is most often open with respect to electron flow. This usage of the term is now much more common among electrical engineers than the original meaning, "the complete path of an electric current including usually the source of electric energy" (Woolf, 1981), which implies a closed system with respect to electron flow. It is no accident that the usage of the word *circuit* has evolved in this manner. Early in the development of electrical technology, a useful system [such as the electromagnetic telegraph (Marland, 1964)] was composed of at most a few topologically closed "circuits," and the closure of the current path was a central concern. As the complexity of electrical systems increased, the convention of organizing a system in terms of a power supply and its buses was developed. This provided a common segment for all the current paths, and the attention of the engineer focused on the remaining, "interesting" segment, that which contained the active devices (and the term *circuit* came to be applied to such a segment). However, by focusing on only a segment of the current path, one had to deal with an open system, rather than a closed one.

The physics of closed systems is certainly simpler than that of open systems, because closed systems obey global conservation laws, while open systems, in general, do not. In the well-established techniques of physical theory one often encounters artifices, usually in the form of periodic boundary conditions, which assure the closure of the theoretical model, if not of the system itself. The point of the present discussion is that it is frequently necessary to partition a complex system (which might reasonably be regarded as closed) into smaller components which, viewed individually, must be regarded as open. Thus, the more applied disciplines of the physical sciences must often deal at some level with the concept of an open system.

There are many established techniques for dealing with open systems in fields such as fluid dynamics, neutron transport, and electronics. All these fields are concerned with the transport of (usually) conserved particles. The transport phenomena are described by transport equations at a kinetic or hydrodynamic level which are either

differential or integro-differential equations. Such equations require boundary conditions, and it is in these boundary conditions that the openness of a system is described. In the computation of the flow around an airfoil, one must supply "upstream" and "downstream" boundary conditions (Roache, 1976, Chap. III, Sec. C). In electronics the connection to the external circuit is accomplished by some sort of contact. In solid-state electronics the most frequently used type of contact is the ohmic contact, an interface between a metallic conductor and (usually) a semiconductor which permits electrons to pass freely. Because the ohmic contact is a critical component of solid-state technology, most work on such interfaces has been directed toward their fabrication and characterization (Milnes and Feucht, 1972). The theoretical representation of such contacts by boundary conditions has been a part of the analysis of semiconductor device problems since the beginning of semiconductor technology (Bardeen, 1949; Shockley, 1949). The current practice of using boundary conditions to model contacts is discussed in detail by Selberherr (1984, Sec. 5.1).

B. Theoretical approaches to open quantum systems

Since the existing theoretical work on open systems consists primarily of the definition of boundary conditions on transport equations, it is appropriate to examine various approaches to transport theory to see how they have dealt with this issue. This examination will center upon electron-transport theory, because we wish to include quantum-coherence effects in the theory, and these are much more prominent in systems of electrons than in systems of more massive particles.

By far the most common approach to defining the boundary conditions on a transport problem is to circumvent the issue entirely. This is most easily done by restricting one's attention to the special case of spatially uniform systems, so that (at the kinetic level) all spatial derivatives disappear, and with them the need to specify the boundary conditions. Applications of the Boltzmann equation (as expressed in terms of the usual Euler variables) have most often been restricted to the case of uniform driving fields (Dresden, 1961; Conwell, 1967). When the Boltzmann equation has been applied to nonuniform systems (see, for example, Castagné, 1985; Constant, 1985; Reggiani, 1985; Baranger and Wilkins, 1987), techniques requiring that the equation be recast in terms of the Lagrange variables have generally been employed. Boundary conditions for such formulations are discussed in Appendix C. Much of the work on quantum transport has also assumed uniform fields (see, for example, Levinson, 1969; Mahan, 1987).

The other popular approach is to assume periodic boundary conditions (Kohn and Luttinger, 1967), which assure the Hermiticity of all relevant operators (Yennie, 1987). This in effect closes the system, forestalling the possibility of studying any open-system aspects of the problem. It also prevents one from studying any situation in which the change in chemical potential across the system is of finite magnitude (because the potential must also be periodic). Periodic boundary conditions are thus adapted to the requirements of linear-response theory (Kubo, 1957), but not to those of far-from-equilibrium problems.

A fundamental approach that does take cognizance of the open nature of transporting systems is that advocated by Landauer (1957, 1970; also Büttiker *et al.*, 1985). This approach envisions a system within which dissipative processes do not occur, but which is coupled to two or more ideal particle reservoirs. The conductance of such a structure is then expressed in terms of the quantum-mechanical transmission coefficients of the system. The ideal reservoirs have properties analogous to those of a blackbody: They absorb without reflection any electrons leaving the system and emit an equilibrium thermal distribution into the system. We shall see that such a picture is indeed the key to constructing a useful open-system model. However, let us note that this approach does not specify the boundary conditions on a boundary-value problem. The boundary conditions are actually applied to Schrödinger's equation and are the asymptotic conditions upon which the formal theory of scattering is based (see Appendix D). The concept of thermal reservoirs is invoked to specify how the various wave functions are to be incorporated into a density operator for the system, from which observables may be evaluated.

The Landauer approach has successfully described a number of quantum conductance phenomena (Stone and Szafer, 1988): Aharonov-Bohm oscillations, universal conductance fluctuations, and quantized conductance through constrictions (Szafer and Stone, 1989). (Many recent results in this area can be found in Heinrich, Bauer, and Kuchar, 1988, and in Reed and Kirk, 1989.) However, it is important to recognize that these phenomena occur only under a very restricted range of circumstances (Webb, 1989): cryogenic temperatures (typically 1 K) and low voltages (typically 1 meV). The reason for this is not so much the fragility of quantum-interference effects in themselves, but rather the constraints placed upon the phenomena by the requirement that they be observable in the linear-response regime (which is to say, very near to thermal equilibrium). Near equilibrium, only the states near the Fermi level contribute to the conductance, but *all* such states participate. As the temperature or bias voltage is raised, more states participate in the conduction, with slightly different energies or wave vectors, and the observable effects are "washed out."

In a far-from-equilibrium situation one has the opportunity to populate selectively a narrow set of quantum states, leaving nearby states unpopulated. This can lead to quantum-interference phenomena which are quantitatively dominant at or above room temperature. The prototypical example of such a situation is provided by the quantum-well resonant-tunneling diode (Chang, Esaki,

and Tsu, 1974; Sollner *et al.*, 1983), which is discussed more extensively in Sec. V. Such devices have demonstrated peak-to-valley current ratios as high as 30 at 300 K (Broekaert, Lee, and Fonstad, 1988; for a tabulation of device results see Mehdi and Haddad, 1989).

Given that far-from-equilibrium quantum-interference effects can be large and are thus important to study, one must ask whether such effects can be adequately described by elementary quantum theory. For the case of tunneling structures the standard elementary theory assumes that the electron states are stationary scattering-state solutions of Schrödinger's equation (Duke, 1969; Tsu and Esaki, 1973; Wolf, 1985). Does this provide an adequate description of nonequilibrium phenomena? The answer is, in general, no, and we shall explore this issue below. The elementary tunneling theory does seem to give good results for the current density, but for other physical observables, such as the charge distribution, a more sophisticated approach is required.

To demonstrate the problems one encounters with elementary quantum-mechanical models in a far-from-equilibrium situation, let us consider the apparently simple problem of finding the self-consistent electrostatic potential in a single-barrier tunneling structure. A semiconductor heterostructure is assumed, and the details of the structure and analysis are given in Appendix A. The approach that we shall use is first to approximate the self-consistent potential using the Thomas-Fermi screening theory. The resulting potential and electron distribution are shown in Fig. 1. The Thomas-Fermi potential shows the smooth bending that one would expect in a system in which the charge densities are several orders of magnitude less than those in metallic systems. Now we use the Thomas-Fermi potential in Schrödinger's equation and start an iterative procedure to find the "true" self-consistent potential. The results of the first iteration are also shown in Fig. 1, and it is quite clear that we will not obtain a physically credible result. The charge density obtained from Schrödinger's equation differs markedly from the Thomas-Fermi solution on the left-hand (upstream) side of the barrier. Where Thomas-Fermi indicates an accumulation of electrons, Schrödinger's equation gives a depletion of electrons. The reason for this is that the tunneling theory assumes that the electron states in the potential "notch" on the left-hand side of the barrier are in equilibrium with the right-hand reservoir, because that is the side from which these wave functions are incident. The depletion of electron density may be traced to the requirement of current continuity in the propagating states: As an electron propagates into a region of decreasing potential, its velocity increases; but to maintain a constant current density, its amplitude must then decrease. Because the tunneling-theory charge density does not produce overall charge neutrality in the structure, the solution of Poisson's equation has large electric fields at the boundaries, which in turn exacerbates the problem of charge neutrality. (The final self-consistent result would show the energy barrier lying near the bottom of a parabolic potential well considerably deeper than that of the first iteration.)

The physical processes that work to enforce charge neutrality are those which work to restore thermal equilibrium, which is to say, inelastic processes. In the present case, these are the inelastic scattering events (primarily phonon scattering) which dissipate the electrons' energy and cause electrons in the propagating states entering from the left-hand reservoir to fall into the lower-energy notch states. The resulting population in the notch states produces the accumulation of negative charge required to screen the electric field. Thus the true self-consistent potential will depend upon the number of electrons in the notch states, which in turn will depend upon the relative rates at which electrons are scattered into the notch states and subsequently tunnel out (Wingreen and Wilkins, 1987). Therefore a physically

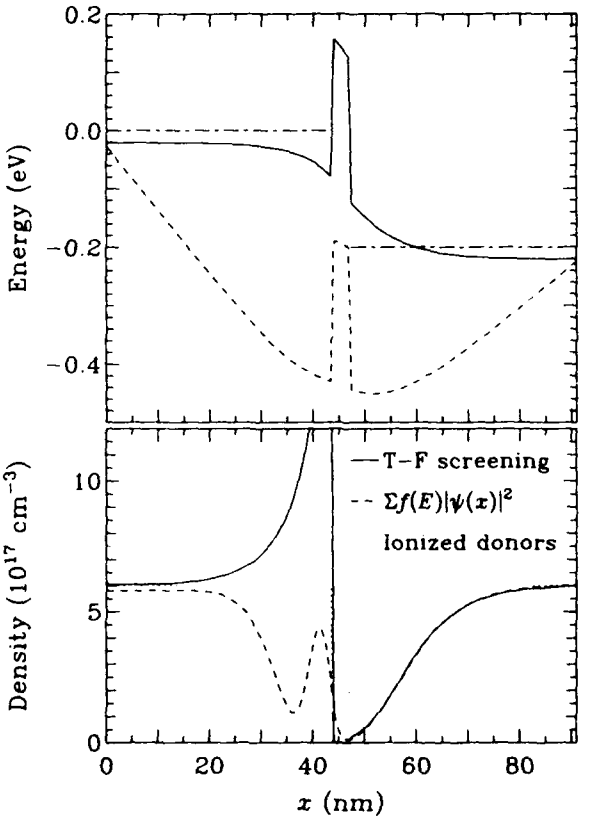


FIG. 1. Potential (upper) and charge density (lower) of a semiconductor tunneling heterostructure biased far from equilibrium: solid lines, results of a Thomas-Fermi screening model; dashed lines, charge density and first iteration of the potential obtained by solving Schrödinger's equation in a conventional tunneling calculation. The tunneling result fails to display an accumulation of electrons on the upstream side of the barrier because inelastic processes are not included, and as a result the self-consistent potential is quite unphysical. The dotted line shows the distribution of positive charges, and the dot-dashed line shows the chemical potentials.

reasonable self-consistent potential will not be obtained unless the inelastic processes are included in the analysis.

The usual way to incorporate inelastic processes, to the first order, is to use the Fermi golden rule to evaluate the transition rates between states. In a more complete description these transition rates actually appear as terms in a Pauli master equation (see Kreuzer, 1981, Chap. 10). The Pauli master equation assumes that the electrons occupy only eigenstates of the Hamiltonian, not superpositions of those eigenstates. In other words, the density operator of the system is and remains diagonal in the eigenbasis of the Hamiltonian. In the present case, this assumption violates continuity. An example of this is shown in Fig. 2, which shows two eigenstates of Schrödinger's equation, one incoming from the left and one confined in the notch (though it is coupled by tunneling to a propagating state on the right-hand side of the barrier). An inelastic process described by the Pauli master equation will cause probability density to disappear from one state and reappear in the other. Because the spatial distributions of the two states are different, this means that the probability distribution must change with time. But because the states are both eigenstates, their current densities are uniform. Thus the Pauli master equation violates the continuity equation. This is explored more formally in Appendix B. Presumably, inelastic transitions are more localized processes, involving superpositions of eigenstates which describe such localization. However, this implies that the off-diagonal elements of the density operator are non-negligible, and theories that comprehend off-diagonal density operators are kinetic theories.

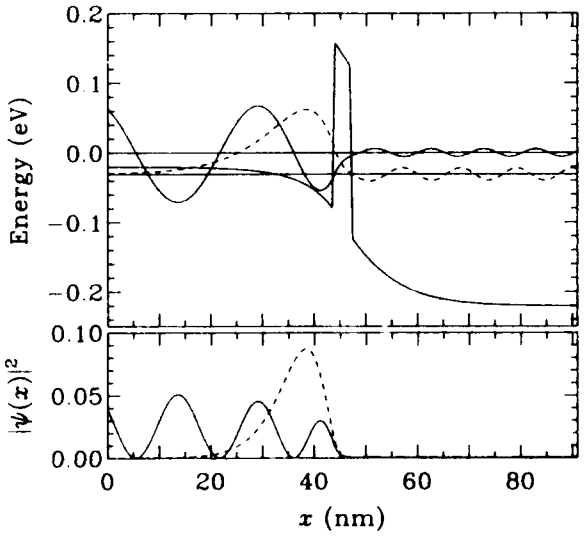


FIG. 2. Typical eigenstates in a tunneling structure: solid line, a propagating state; dashed line, a state that is confined in the potential "notch." The spatial distributions of these states are quite different, as shown in the lower plot of $|\psi(x)|^2$. Thus the Pauli master-equation description of an inelastic process that couples these states must violate the continuity equation.

To demonstrate that a plausible solution to the self-consistent-potential problem can be obtained using kinetic theory, the results of such a calculation are shown in Fig. 3. The approach described in Secs. IV and V was used, and inelastic processes (phonon scattering) were included using the Boltzmann collision operator described in Appendix F. When the phonon scattering processes are included [Fig. 3(a)], an accumulation layer is formed in the potential notch. However, the accumulation is not sufficient to screen the electric field effectively as it approaches the boundary. Evidently there are other effects that need to be included. One such effect is the resistivity of the contacting layers (outside of the calculation domain). If these layers are ohmic conductors, the distribution of electrons in them must shift away from its equilibrium value when a current is conducted. When this effect is incorporated into the boundary conditions on the kinetic model, the self-consistent potential shown in Fig. 3(b) is obtained. This is a much more credible result, as

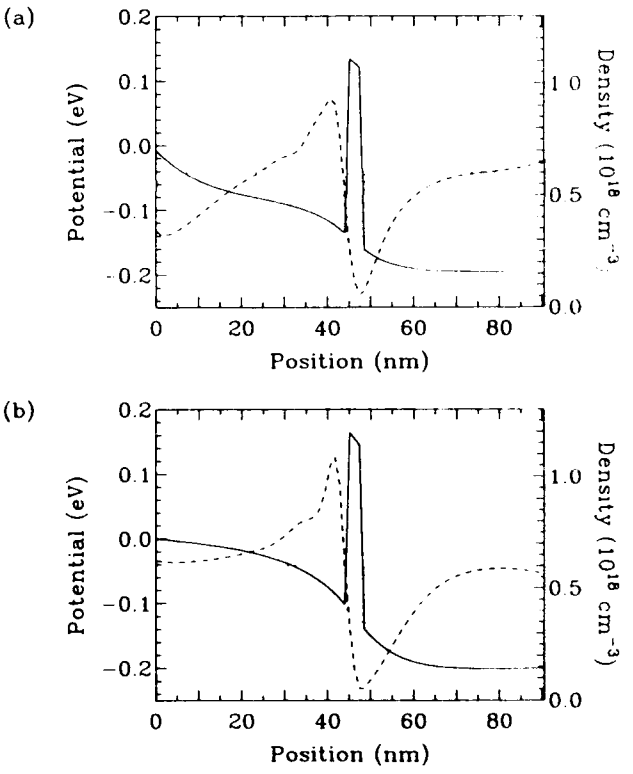


FIG. 3. Calculations of the self-consistent potential of the tunneling heterostructure using a kinetic theory that includes inelastic processes. In (a) the longitudinal-optic and acoustic-phonon scattering processes are included, but the incoming distribution of electrons is fixed. An accumulation layer is formed on the upstream side of the barrier, but the screening of the electric field is far from complete. In (b) the incoming distribution of electrons is allowed to shift in response to the electric field at the boundary, to simulate an Ohmic conductor outside the boundary. The screening is more complete, and the resulting potential is more physically credible.

the potential varies smoothly through the structure and the electric field approaches a small value at the boundaries. The screening length from the kinetic model is significantly longer than the value indicated by the Thomas-Fermi calculation of Fig. 1. This might have been expected from the effects of size quantization in the notch (Ando, Fowler, and Stern, 1982) and also from the finite rate of inelastic transitions that fill the notch.

Thus the problem of calculating the self-consistent potential in a tunneling structure is about as complicated as it could possibly be, in the sense that the qualitative result depends upon all the processes occurring within the system. It thus provides a vivid example of the problems one encounters in attempting to apply elementary quantum-mechanical concepts to a far-from-equilibrium situation. A satisfactory treatment of far-from-equilibrium phenomena requires an approach at a level of sophistication at least equal to that of kinetic theory.

II. QUANTUM KINETIC THEORY

A. Levels of approximation in statistical theory

A generally accepted approach to the problems of statistical physics is to begin with the general theory of many-body dynamics and to proceed by deductive reasoning to a formulation that provides an answer for the problem of interest (see, for example, Reichl, 1980). The steps in this deductive chain necessarily involve the introduction of extra assumptions in the form of suitable approximations. One may loosely categorize the levels of approximation in terms of the independent variables required to specify the state of a system. The most detailed level is the fundamental many-body theory, which in principle requires a complete set of dynamical variables for each particle. This can be reduced to the kinetic level by restricting one's attention to one- or two-body properties [by truncating the BBGKY hierarchy of equations, for example (Reichl, 1980, Sec. 7C)]. It may also be necessary to remove from explicit consideration other dynamical variables of the complete system, such as photon or phonon coordinates, when electrons are the particles of interest. The kinetic theory is expressed in terms of distribution functions defined on a single-particle phase space, requiring one position and one momentum variable for each spatial dimension. (In the quantum case, this goes over to two arguments of the density operator.) The hydrodynamic level of approximation is obtained by making some assumption about the form of the distribution function with respect to momentum, and integrating over all momenta. Thus the hydrodynamic theory is expressed in terms of densities that are functions of position only.

The approach taken in the present work is quite different from the conventional deductive approach. The objective is to identify the mathematical properties required of simple kinetic models of open systems. The procedure will be to construct small, spatially discretized models and to explore their properties numerically. The significance of the results must then be argued inductively.

B. Fundamentals of kinetic models

In the kinetic level of description of a complex system, the effects of those degrees of freedom that are of less interest in a given problem are included implicitly in objects such as collision operators or effective interaction potentials. In the example of electronic devices such degrees of freedom should include electron coordinates outside the device, but within the external circuit. They also include all excitations of the device material apart from the single-electron states (e.g., the phonons). Thus, at this level, the state of the system is described by a one-body density operator or distribution function. In general, this can be written as

$$\rho(x,x') = \sum_i w_i \langle x\, i \rangle \langle i\, x' \rangle , \tag{2.1}$$

where i labels a complete set of states and the w_i are real-valued probabilities for the system to be in state $|i\rangle$. Because we shall be considering open systems in which the number of particles is not fixed, the usual convention for the normalization of $\rho(\langle i\, i \rangle = 1$ and $\mathrm{Tr}\rho = 1)$ is not useful. Instead, we shall adopt a normalization convention such that $\rho(x,x)$ gives the actual particle density (in units of particles per cm^3, for example). More formally, ρ is the one-body reduced density operator which is defined on a single-particle Hilbert space (Reichl, 1980, Chap. 7). The complete density matrix defined on the many-particle Fock space (second quantization) may still be normalized to unity. The focus upon a single-particle description requires that one exercise some care concerning the quantum statistics. For example, if the equilibrium density operator is obtained by solving the Bloch equation, $\partial\rho/\partial\beta = -H\rho$, the result will satisfy Maxwell-Boltzmann statistics. A similar calculation in the Fock space will, of course, satisfy Fermi-Dirac statistics.

For a system described by a simple single-particle Hamiltonian,

$$H = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + v(x) , \tag{2.2}$$

the time evolution of the density matrix is given by the Liouville-von Neumann equation:

$$i\hbar\frac{\partial\rho}{\partial t} = [H,\rho] = \mathcal{L}\rho$$
$$= -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right)\rho + [v(x) - v(x')]\rho , \tag{2.3}$$

where $\mathcal{L}$ is the Liouville superoperator. The simplest approach to modeling the behavior of open systems is to ap-

ply the Liouville equation to a finite spatial domain representing the system of interest and to apply boundary conditions that model the openness of the system. The difficulties and ultimate success of this approach involve the effect that such boundary conditions have upon the properties (particularly the eigenvalue spectrum) of the Liouville superoperator.

C. Linear algebra of superoperators

A central issue in the development of a kinetic model for open systems is the stability of the resulting time-dependent solutions, which depends upon the eigenvalue spectrum of the Liouville superoperator. Zwanzig (1964) has presented an excellent discussion of the properties of superoperators (or tetradics). However, the present analysis requires a somewhat different group of expressions, so the subject will be developed here. The density operators that represent the state of a statistically mixed system themselves form a linear vector space analogous to the space of pure quantum states represented by wave functions. A linear combination of density operators might be used to describe the results of superposing two partially polarized beams of particles, for example (using the present normalization of ρ). Anything that generates linear transformations on a density operator [such as the right-hand side of the Liouville equation (2.3)] is a superoperator. In a finite, discrete system with N states, a wave function will be a vector (a singly-indexed object) with N elements, the density operator will be a matrix (a doubly-indexed object) with N^2 elements, and a superoperator will be a tetradic (a quadruply-indexed object) with N^4 elements. The linear algebra of superoperators is isomorphic to that of ordinary operators, but to define concepts such as Hermiticity or unitarity of superoperators, we must have a definition for the inner product of two ordinary operators. The simplest definition is

$$\langle A \| B \rangle = \mathrm{Tr}(A^\dagger B) , \tag{2.4}$$

where A and B are operators and the notation $\langle \| \rangle$ is introduced to indicate expressions in the linear space of operators. It is easily shown that this satisfies the axioms (Apostol, 1969) defining an inner product on a complex vector space. Then a Hermitian superoperator $\mathcal{H}$ satisfies

$$\langle A \| \mathcal{H} B \rangle = \langle \mathcal{H} A \| B \rangle , \tag{2.5}$$

and a unitary superoperator $\mathcal{U}$ satisfies

$$\langle \mathcal{U} A \| \mathcal{U} B \rangle = \langle A \| B \rangle . \tag{2.6}$$

Superoperators are usually derived from ordinary quantum observable operators by forming the commutator or anticommutator with the operator being acted upon. For an operator C let us denote these superoperators

$$\mathcal{C}_{(-)} A = CA - AC , \tag{2.7}$$

$$\mathcal{C}_{(+)} A = \tfrac{1}{2}[CA + AC] . \tag{2.8}$$

If C is Hermitian ($C^\dagger = C$), the Hermiticity of $\mathcal{C}_{(-)}$ and $\mathcal{C}_{(+)}$ follow immediately:

$$\begin{aligned}\langle \mathcal{C}_{(-)} A \| B \rangle &= \mathrm{Tr}[(CA - AC)^\dagger B] = \mathrm{Tr}(A^\dagger C^\dagger B - C^\dagger A^\dagger B) \\ &= \mathrm{Tr}(A^\dagger C B - C A^\dagger B) = \mathrm{Tr}(A^\dagger C B - A^\dagger B C) \\ &= \mathrm{Tr}[A^\dagger (CB - BC)] = \langle A \| \mathcal{C}_{(-)} B \rangle ,\end{aligned}$$

and similarly for $\mathcal{C}_{(+)}$. The Hermiticity (or lack thereof) of the Liouville superoperator is the critical issue in formulating a kinetic model of open systems.

Of particular importance are the superoperators generated by the position operator x and the momentum operator $p_x = (\hbar/i)\partial/\partial x$:

$$\mathcal{X}_{(+)} = \tfrac{1}{2}(x + x') , \tag{2.9}$$

$$\mathcal{X}_{(-)} = x - x' , \tag{2.10}$$

$$\mathcal{P}_{(+)} = \frac{\hbar}{2i}\left[\frac{\partial}{\partial x} - \frac{\partial}{\partial x'}\right] , \tag{2.11}$$

$$\mathcal{P}_{(-)} = \frac{\hbar}{i}\left[\frac{\partial}{\partial x} + \frac{\partial}{\partial x'}\right] . \tag{2.12}$$

These superoperators obey the following commutation relations:

$$[\mathcal{X}_{(+)}, \mathcal{P}_{(+)}] = [\mathcal{X}_{(-)}, \mathcal{P}_{(-)}] = 0 , \tag{2.13}$$

$$[\mathcal{X}_{(+)}, \mathcal{P}_{(-)}] = [\mathcal{X}_{(-)}, \mathcal{P}_{(+)}] = i\hbar . \tag{2.14}$$

Thus $\mathcal{X}_{(+)}$ is in some sense conjugate to $\mathcal{P}_{(-)}$, and $\mathcal{X}_{(-)}$ bears a similar relationship to $\mathcal{P}_{(+)}$. Of course $\mathcal{C}_{(+)}$ commutes with $C_{(-)}$ for any operator C.

D. Irreversibility

Kinetic theory appears to be the simplest level at which one may consistently describe both quantum interference and irreversible phenomena (Prigogine, 1980). The only available levels that are simpler, in that they require fewer independent variables, are hydrodynamics and elementary (single-particle, pure-state) quantum mechanics. Hydrodynamics (as embodied in Ohm's law and the drift-diffusion equation in solid-state physics) provides no means to describe quantum effects such as resonance phenomena because it retains no information on the distribution of particles with respect to energy or momentum. On the other hand, if one attempts to include irreversible processes within the framework of elementary quantum mechanics, the continuity equation is most often violated. Irreversible processes will generally result in the time dependence of some physical observable showing an exponential decay. The only time dependence provided by elementary quantum theory is the $e^{-iEt/\hbar}$ dependence of the wave function. Exponential decay implies that E must have a negative imaginary part, which means that the electron (for example) ex-

ponentially disappears, violating charge conservation. As we have seen, violations of continuity still occur when the irreversible processes are described by the Fermi golden rule or Pauli master equation (see Appendix B). To maintain consistency with the continuity equation, we must allow off-diagonal elements of the density matrix (in the eigenbasis of the Hamiltonian) to develop as the system evolves (see Peierls, 1974). Because we do not know *a priori* which off-diagonal elements are required, we must admit all off-diagonal elements. A theory that describes the evolution of the complete (single-particle) density operator, including the off-diagonal elements, is by definition a kinetic theory.

To express this point in another way, *we cannot, in general, assume that the particles in an irreversible system occupy the eigenstates of the Hamiltonian.* The proper basis states for a one-particle description are the eigenstates of the density operator, and thus the specification of the basis set should be a result obtained from a proper theory, rather than an *a priori* assumption in the theory. The exception to this situation is the particular case of thermal equilibrium. In this case we know that the density operator is a function of the Hamiltonian (via the Bloch equation, $\rho \propto e^{-\beta H}$), and if an effective one-particle Hamiltonian is an adequate description, the particles in the system will be found in eigenstates of this Hamiltonian, if they are in equilibrium.

The usual way to describe the effects of irreversible or dissipative processes at the kinetic level is to add a collision term (of one form or another) to the Liouville equation (2.3) to obtain a Boltzmann equation. This is a valid procedure so long as the dissipative processes are sufficiently weak that the motion of the particles can be viewed as periods of free flight interrupted by collision events. Such a term takes its simplest form for interactions between the particles of interest (i.e., electrons) with particles that either are spatially fixed (such as impurities in solids) or can be modeled as components of a thermal reservoir (such as the phonons). In this case (and within the Markov assumption) the collision term is a simple linear superoperator expression, and we can write the Boltzmann equation as

$$\partial\rho/\partial t = (\mathcal{L}/i\hbar)\rho + \mathcal{C}\rho \ , \tag{2.15}$$

where $\mathcal{C}$ is the collision superoperator. (We shall see later what condition $\mathcal{C}$ must satisfy to preserve the continuity equation.) For two-body collisions the operator is a more complex object, operating on a two-body density matrix or (if the Stosszahlansatz is invoked) a product of one-body density matrices which introduces nonlinearity.

A characteristic feature of irreversible systems is the existence of stable stationary states, which can be either the equilibrium state or a nonequilibrium steady state if the system is driven by an external agency. Perturbations upon such a steady state will, in general, decay. To describe this decay the Boltzmann superoperator $\mathcal{L}/i\hbar + \mathcal{C}$ must have eigenvalues with negative real parts. In the usually studied case the Liouville superoperator is Hermitian, so $\mathcal{L}/i\hbar$ by itself would produce purely imaginary eigenvalues. The collision operator $\mathcal{C}$ introduces the negative real parts of the eigenvalues. Physically, we expect that there should be no eigenvalues with positive real parts, because these would correspond to exponentially growing modes, and the system would not be stable. The presence of eigenvalues with negative real parts together with the absence of eigenvalues with positive real parts implies that the system is time irreversible.

The study of the fundamental origins of irreversibility in physical theory remains an area of active discussion and debate, more than a century after the question was first raised. However, if one's objective is to develop useful models of physical systems with many dynamical variables, rather than to construct a rigorously deductive mathematical system, it is clearly most profitable to adopt the view that irreversibility is a fundamental law of nature. For the present purposes a more precise statement of this law is that "simple" systems will always stably approach a steady state. In this context simple systems are those which can be regarded as being composed of a single type of particle or single chemical species and such that all other types of particle or excitation can be represented by thermal reservoirs. [Multicomponent systems can display exponential growth or stable oscillation (Prigogine, 1980).] The stability of the physical system implies that the kinetic superoperator that generates the time evolution of the density matrix (whether it be of the Liouville, Boltzmann, or some other form) cannot possess eigenvalues that would lead to growing exponential solutions. That is, there can be no eigenvalues with a positive real part. This condition will determine the sort of boundary conditions that can be used to model open systems.

Throughout most of the present analysis the collision terms will be neglected, because we shall see that irreversibility enters through the open-system boundary conditions. The irreversible open-system model permits a wide variety of phenomena to be described at least qualitatively without invoking a collision term. This is not to say that irreversible collisions or dissipative interactions within a system are not significant effects. Indeed, a central thrust of traditional transport theory is the derivation of kinetic descriptions of such phenomena. The present neglect of the collision term is merely for the sake of simplicity, and it should be borne in mind that such a term may be readily added to any of the calculations to be discussed (see Appendix F).

III. TIME-REVERSIBLE OPEN-SYSTEM MODEL

To describe the behavior of an open system, we shall consider an approach in which the spatial domain is considered to be finite, corresponding to the extent of the system, and boundary conditions are applied which permit particles to pass into and out of the system. The first model we shall consider employs time-reversible boundary conditions which are plausible, but which we shall ul-

timately see to be unphysical (Frensley, 1985). This model helps to define the conditions that a physically reasonable open-system model must display.

A. Continuum formulation

To provide the motivation for the first model, let us consider a spatially uniform particle gas of infinite extent, $-\infty < x < \infty$, and take the open system to be the finite region $0 \le x \le l$. The thermal equilibrium density matrix for a uniform gas may be obtained by integrating the Bloch equation (Feynman, 1972)

$$\partial\rho_{\rm eq}/\partial\beta = -H\rho_{\rm eq} . \tag{3.1}$$

The solution $\rho_{\rm feq}$ (for free particles in equilibrium) is

$$\rho_{\rm feq}(x,x') = \frac{1}{\sqrt{2\pi}\lambda_T}\exp[-(x-x')^2/2\lambda_T^2 + \beta\mu] , \tag{3.2}$$

where the normalization is such that $\rho_{\rm feq}(x,x)$ gives the number of particles per unit length, μ is the chemical potential, and λ_T is a thermal coherence length given by

$$\lambda_T^2 = \hbar^2\beta/m . \tag{3.3}$$

Now if we arbitrarily impose boundaries along the lines $x=0$, $x=l$, $x'=0$, and $x'=l$, what boundary conditions would $\rho_{\rm feq}$ satisfy? Note that the dependence is only upon $(x-x')$, so that $\partial\rho/\partial x = -\partial\rho/\partial x'$. Thus in this particular case ρ obeys the homogeneous boundary condition

$$\left[\frac{\partial}{\partial x} + \frac{\partial}{\partial x'}\right]\rho\Big|_{\rm boundary} = 0 . \tag{3.4}$$

In other words, the directional derivative of ρ in a direction parallel to the principal diagonal is set to zero at the boundaries.

Is Eq. (3.4) the appropriate boundary condition for a general open system? Let us explore some of its consequences. Suppose at time $t=0$ we apply a uniform force field F to the particle gas. The solution to the Liouville equation (2.3) over the infinite domain and with initial condition (3.2) describes an accelerating gas and is given by

$$\rho_{\rm acc}(x,x';t) = \rho_{\rm feq}(x,x')\exp\left[\left[\frac{iFt}{\hbar}\right](x-x')\right] . \tag{3.5}$$

Now $\rho_{\rm acc}$ also obeys Eq. (3.4), so it is also the solution to Eq. (2.3) over the finite domain subject to boundary condition (3.4).

A more general consequence of boundary condition (3.4) is that the particle densities at the boundaries, $\rho(0,0)$ and $\rho(l,l)$, remain constant as the density matrix evolves with time. To demonstrate this, note that we can factor the hyperbolic operator in the Liouville equation (2.3) derived from the kinetic energy terms as

$$-\frac{\hbar^2}{2m}\left[\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial x'^2}\right] = -\frac{\hbar^2}{2m}\left[\frac{\partial}{\partial x} - \frac{\partial}{\partial x'}\right]\left[\frac{\partial}{\partial x} + \frac{\partial}{\partial x'}\right] = \frac{1}{m}\mathcal{P}_{-}\mathcal{P}_{+} . \tag{3.6}$$

The boundary condition assures that the second factor in Eq. (3.6) is zero along the boundaries, and along the diagonal the potential term is zero. Thus $\partial\rho(0,0)/\partial t = 0$ and $\partial\rho(l,l)/\partial t = 0$. This might be interpreted as the behavior of a large reservoir with a fixed particle density (or fixed pressure if the temperature is also fixed). Thus the boundary condition (3.4) provides a plausible model for an open system.

In fact, the Liouville equation (2.3) subject to the boundary condition (3.4) generates an unphysical solution in the form of exponentially growing particle densities when it is applied to more general potentials that do not have the symmetry of the uniform field (Frensley, 1985). The nature of the time-dependent solutions (whether they be growing, decaying, or oscillating) depends upon the eigenvalue spectrum of the Liouville superoperator (the definition of which requires both the differential operator and the boundary conditions). The problem with the growing densities (and ultimately the identification of the correct model) is a consequence of opening the system, which violates the Hermiticity of the Hamiltonian operator and of the Liouville superoperator. Recall the proof (Messiah, 1962) of the Hermiticity of the Hamiltonian (2.2). It proceeds by invoking Green's identity to transpose the Laplace operator, which leaves a surface term. The precise expression is

$$\int_\Omega (H - H^+)d^3\mathbf{x} = \frac{\hbar}{i}\int_S \mathbf{j}\cdot d^2\mathbf{s} , \tag{3.7}$$

where Ω refers to the volume of the domain, S is its surface, and $\mathbf{j}$ is the current-density operator. One maintains the Hermiticity of the Hamiltonian by choosing basis functions for which the surface integral is identically zero: states well localized within the domain and stationary scattering states (or periodic boundary conditions) for which the incoming and outgoing currents cancel. Because the total number of particles in an open system can change in response to externally imposed conditions, such a basis set is too restrictive.

The violation of the Hermiticity of the Liouville superoperator follows directly from the violation of Hermiticity of the Hamiltonian. This leads to eigenvalues of the Liouville superoperator that have nonzero imaginary parts, leading to real exponential behavior in the time dependence of ρ. As mentioned previously, the inclusion of dissipative interactions will introduce decaying exponential behavior. It is thus quite enlightening to observe both the separate and combined effects of dissipa-

tion and open-system boundary conditions on the eigenvalue spectrum of the Liouville superoperator (though technically it is no longer the Liouville operator when dissipation is included). For this purpose let us consider an extremely simple model of dissipation. This model is simple Brownian motion as described by the Fokker-Planck or Kramers equation (Kubo, Toda, and Hashitsume, 1985). It is classically valid in the limit that the particles of interest are weakly coupled to an ideal reservoir. Caldeira and Leggett (1983) have studied the quantum-mechanical derivation of this equation and have shown it to be valid at higher temperatures ($\hbar\beta$ smaller than or comparable to the response time of the reservoir to which the particles are coupled). In terms of ρ the Fokker-Planck equation may be written in the form of Eq. (2.15) with the collision operator given by

$$\mathcal{C}_{\rm FP}\rho = -\gamma\left[\frac{(x-x')}{2}\left[\frac{\partial}{\partial x}-\frac{\partial}{\partial x'}\right]\rho+\frac{m}{\hbar^2\beta}(x-x')^2\rho\right]$$
$$= -\gamma(i\mathcal{X}_{(-)}\mathcal{P}_{(+)}/\hbar+\mathcal{X}^2_{(-)}/\lambda_T^2)\rho\ , \qquad (3.8)$$

where γ is the damping rate. The first term in Eq. (3.8) describes dissipation and corresponds to a frictional force equal to γp, where p is the linear momentum. The second term describes the thermal fluctuations. An important property of $\mathcal{C}_{\rm FP}$ is that $(\mathcal{C}_{\rm FP}\rho)(x,x)=0$, which is required for consistency with the continuity equation. $\mathcal{C}_{\rm FP}$ will be used below to add dissipative interactions to our open-system models.

B. Discrete numerical model

To explore the eigenvalue spectrum of the present open-system model and those which will be investigated later, let us consider a finite-difference approximation to the Liouville equation (2.3) which reduces $\mathcal{L}$ to a finite matrix whose eigenvalues may be readily computed. Let me emphasize that only the spatial coordinates will be discretized; time remains continuous, so that the partial differential (and eventually integro-differential) Liouville equation will be reduced to a set of coupled ordinary differential equations with respect to time.

This particular situation requires some discussion. Throughout the computational physics literature, discussions of stability always involve a discretization with respect to time. Because one is accustomed to dealing with continuum equations whose behavior is known to be stable (or at least physical), the common assumption that any instability must be a result of the discretization scheme is generally correct. However, a different situation is being studied here. The validity of the equations themselves (or more precisely the boundary conditions) is the issue. If a discrete-space, continuous-time model is unstable, there will be no time discretization that will correct this instability. On the other hand, we wish to assume that the stability of the discrete-space, continuous-time model will be indicative of the stability of a continuous-space, continuous-time model. As mentioned before, this connection requires a logical induction.

The position coordinates x will be taken to be elements of a uniformly spaced mesh: $\{x_j | x_j = j\Delta_x$ for $j=1,2,\ldots,N\}$. The dependent quantities such as the wave function and density matrix then take on discrete values also, which will be denoted by $\psi_j=\psi(x_j)$ and $\rho_{ij}=\rho(x_i,x_j)$. Using the simple finite-difference approximation $(\partial^2\psi/\partial x^2)_i=(\psi_{i-1}-2\psi_i+\psi_{i+1})/\Delta_x^2$, we find that the Hamiltonian (2.2) becomes

$$H_{ij}=\frac{\hbar^2}{2m\Delta_x^2}(2\delta_{ij}-\delta_{i-1,j}-\delta_{i+1,j})+v_i\delta_{ij}\ , \qquad (3.9)$$

for i,j not on one of the boundaries. To incorporate the boundary conditions, it is best to think of adding an additional mesh point at each end of the domain (points x_0 and x_{N+1}), and specifying the value of the wave function on those points. For example, to apply the homogeneous Dirichlet conditions for a particle in a box, we would set $\psi_0=0$ and $\psi_{N+1}=0$. Inserting these conditions into Eq. (3.9) completely defines the matrix H_{ij} for $1\le i,j\le N$. Similarly, if we wanted to apply Neumann conditions, $\partial\psi/\partial x=0$, we would set $\psi_0=\psi_1$.

Writing the Liouville equation (2.3) on the finite-difference basis gives

$$i\hbar(\partial\rho/\partial t)_{ij}=\mathcal{L}_{ij;kl}\rho_{kl}\ , \qquad (3.10)$$

where the tetradic nature of $\mathcal{L}$ is made explicit. The discrete representation of $\mathcal{L}$ may be derived from Eq. (3.9) and is

$$\mathcal{L}_{ij;kl}=\frac{\hbar^2}{2m\Delta_x^2}(-\delta_{i-1,k}\delta_{jl}-\delta_{i+1,k}\delta_{jl}+\delta_{ik}\delta_{j-1,l}$$
$$+\delta_{ik}\delta_{j+1,l})+(v_i-v_j)\delta_{ik}\delta_{jl}\ . \qquad (3.11)$$

Again, the elements adjacent to a boundary require special attention.

To evaluate the eigenvalues of $\mathcal{L}$ and other superoperators, we must map the tetradic onto an ordinary matrix, so that conventional eigenvalue algorithms may be applied. To do so for the finite, discrete case, we may map the density matrix ρ onto a singly subscripted vector of dimension N^2 by $\rho_{ij}\to\rho_m$ with $m=(i-1)N+j$. Note that with this mapping the inner product between two operators (2.4) becomes the ordinary inner product between two vectors. The mapping of the tetradic $\mathcal{L}$ onto an $N^2\times N^2$ matrix follows immediately. The matrix representing $\mathcal{L}$ was actually constructed for $N=8$ (resulting in a 64×64 matrix for $\mathcal{L}$) using the potential illustrated in Fig. 4. Let us first consider a closed system with no damping. This model is obtained by simply applying the particle-in-a-box (homogeneous Dirichlet) boundary conditions to the Liouville operator (3.11). The resulting eigenvalue spectrum is shown in Fig. 5(a). All the eigenvalues are purely real, as expected from a Hermitian superoperator.

In the second case the model system is taken to be closed, but damped. The Fokker-Planck damping operator (3.8) may be written in discretized form as

$$\mathcal{C}_{ij;kl}=-\frac{\gamma\Delta_x^2}{\lambda_T^2}(i-j)^2\delta_{ik}\delta_{jl}-\frac{\gamma\Delta_x}{2}\times\begin{cases}(i-j)(2\delta_{ik}\delta_{jl}-\delta_{i-1,l}\delta_{jk}-\delta_{ik}\delta_{j+1,l}) & \text{for } i>j\\ (j-i)(2\delta_{ik}\delta_{jl}-\delta_{i+1,l}\delta_{jk}-\delta_{ik}\delta_{j-1,l}) & \text{for } i<j\ .\end{cases} \tag{3.12}$$

This form preserves the important properties of $\mathcal{C}_{\rm FP}$. To illustrate the effect of dissipation on the spectrum of $(\mathcal{L}+i\hbar\mathcal{C}_{\rm FP})$, the zero-temperature limit $(\beta\to\infty)$ was taken (so that the first term, describing fluctuations, vanishes) and the damping constant $\gamma=0.01\omega_0$ (where $\omega_0=\hbar/2m\Delta_x^2$) was used. The resulting eigenvalue spectrum is shown in Fig. 5(b). Negative imaginary parts have been introduced into all the eigenvalues (except possibly one eigenvalue which is equal to zero within the numerical roundoff error, which presumably represents the ground state). These negative imaginary parts lead to damped motion, as expected.

Now with this background we can consider the case of the open-system boundary conditions (3.4) (zero diagonal gradient). The simplest finite-difference approximation for the condition (3.4) is

$$\left|\frac{\partial\rho}{\partial x}+\frac{\partial\rho}{\partial x'}\right|_{ij}=\left|\frac{1}{\Delta_x}(\rho_{i+1,j}-\rho_{ij})+\frac{1}{\Delta_x}(\rho_{ii}-\rho_{i,j-1})\right|$$
$$=\frac{1}{\Delta_x}(\rho_{i+1,j}-\rho_{i,j-1})=0\ , \tag{3.13}$$

for i or j equal to 1 or N. Thus the open-system Liouville superoperator $\mathcal{L}^{(\rm or)}$ (for open system, reversible) is obtained by inserting boundary values $\rho_{i0}=\rho_{i+1,1}$ and $\rho_{N+1,j}=\rho_{N,j-1}$ (and the expressions obtained by transposing the indices) into Eq. (3.11). For the sake of completeness, let us write down the elements of $\mathcal{L}^{(\rm or)}$ that are affected by the boundary conditions:

$$\mathcal{L}^{(\rm or)}_{i1;kl}=\frac{\hbar^2}{2m\Delta_x^2}(-\delta_{i-1,k}\delta_{1,l}+\delta_{ik}\delta_{2,l})+(v_i-v_1)\delta_{ik}\delta_{1,l}\ .$$
$$\mathcal{L}^{(\rm or)}_{1j;kl}=\frac{\hbar^2}{2m\Delta_x^2}(-\delta_{2,k}\delta_{jl}+\delta_{1,k}\delta_{j-1,l})+(v_1-v_j)\delta_{1,k}\delta_{jl}\ . \tag{3.14}$$
$$\mathcal{L}^{(\rm or)}_{iN;kl}=\frac{\hbar^2}{2m\Delta_x^2}(-\delta_{i+1,k}\delta_{Nl}+\delta_{ik}\delta_{N-1,k})$$
$$+(v_i-v_N)\delta_{ik}\delta_{Nl}\ .$$
$$\mathcal{L}^{(\rm or)}_{Nj;kl}=\frac{\hbar^2}{2m\Delta_x^2}(-\delta_{N-1,k}\delta_{jl}+\delta_{Nk}\delta_{j+1,l})$$
$$+(v_N-v_j)\delta_{Nk}\delta_{jl}\ .$$

The non-Hermiticity of $\mathcal{L}^{(\rm or)}$ follows from these expressions. For example, $\mathcal{L}^{(\rm or)}_{i,1;i-1,1}=-\hbar^2/(2m\Delta_x^2)$, but $\mathcal{L}^{(\rm or)}_{i-1,1;i,1}=0$. The boundary conditions have caused elements of $\mathcal{L}$ to be canceled in a way that breaks the Hermitian symmetry. The resulting eigenvalue spectrum is plotted in Fig. 6(a). The non-Hermiticity of $\mathcal{L}^{(\rm or)}$ leads to some eigenvalues with nonzero imaginary parts. It is apparent that these eigenvalues occur in complex-conjugate pairs, with both positive and negative imaginary parts present. This is a consequence of the time-reversal sym-

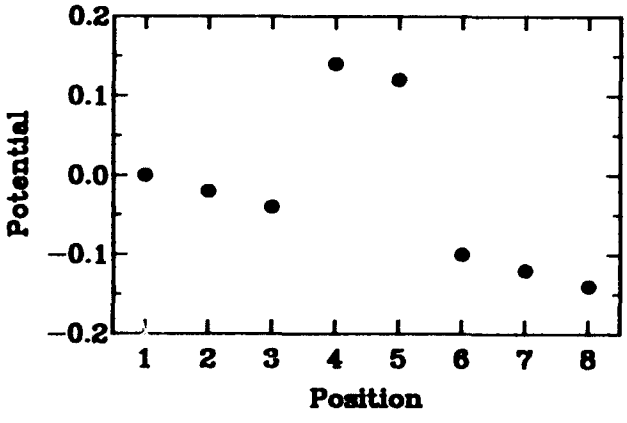


FIG. 4. Potential used in evaluating eigenvalue spectra of Liouville superoperators in the discrete model. This potential was chosen to have both a driving field and a barrier.

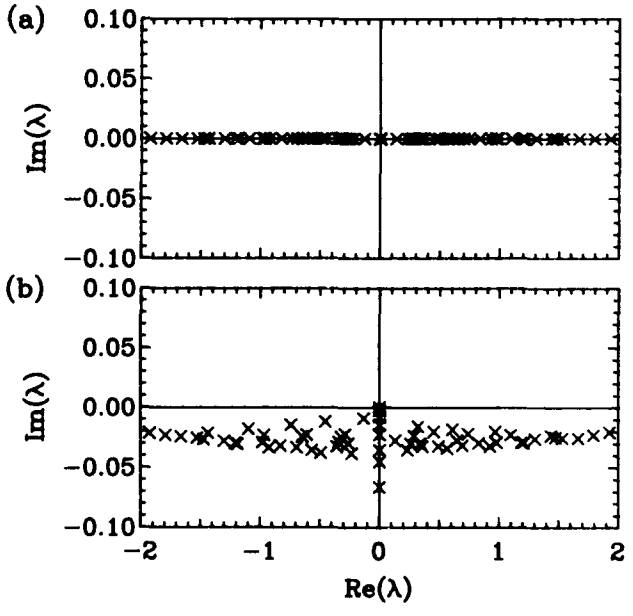


FIG. 5. Eigenvalue spectra of the Liouville operator for a small model closed system with the potential shown in Fig. 4. If the system is taken to be conservative, the resulting eigenvalue spectrum is shown in (a). All eigenvalues are purely real, as expected. In (b) a damping term has been added, leading to negative imaginary parts for most eigenvalues.

metry of both the Liouville equation and the open-system boundary conditions (3.4). The eigenvalues with positive imaginary parts produce growing exponential solutions to the Liouville equation, which would prevent any approach to steady state. This open-system model is thus physically unacceptable.

One might speculate that the problem of growing solutions could be due to the absence of damping in the model. To test this, let us add in the Fokker-Planck damping term (3.12), as we did for the closed-system model. With the same damping constant ($\gamma=0.01\omega_0$) as before, the resulting eigenvalue spectrum for ($\mathcal{L}^{(or)}+i\hbar\mathcal{C}_{FP}$) is that shown in Fig. 6(b). The addition of damping clearly does not solve the stability problem because it does not remove the positive imaginary parts. In fact, a larger damping constant does lead to a stable model, as shown in Fig. 6(c), where $\gamma=0.03\omega_0$ was used. All the eigenvalues now have negative imaginary parts, except for a doubly degenerate eigenvalue at zero (which must be present because of the invariance of ρ_{11} and ρ_{NN}).

Thus modeling an open system by applying the boundary conditions (3.4) will work only if the rate of damping within the system is sufficiently large (or, for the case of electron transport, if the mobility is sufficiently low). The minimum acceptable damping rate depends upon the magnitude of the imaginary parts of the eigenvalues of $\mathcal{L}^{(or)}$ for the undamped system, which in turn depends upon the form of the potential. In fact, the potential of Fig. 4 was chosen because it produces larger imaginary parts than potentials with greater symmetry. All this adds up to a very unsatisfactory formulation for an open-system model. The problems may be traced to the time-reversal symmetry of the boundary conditions. To obtain a proper formulation, this symmetry must be broken.

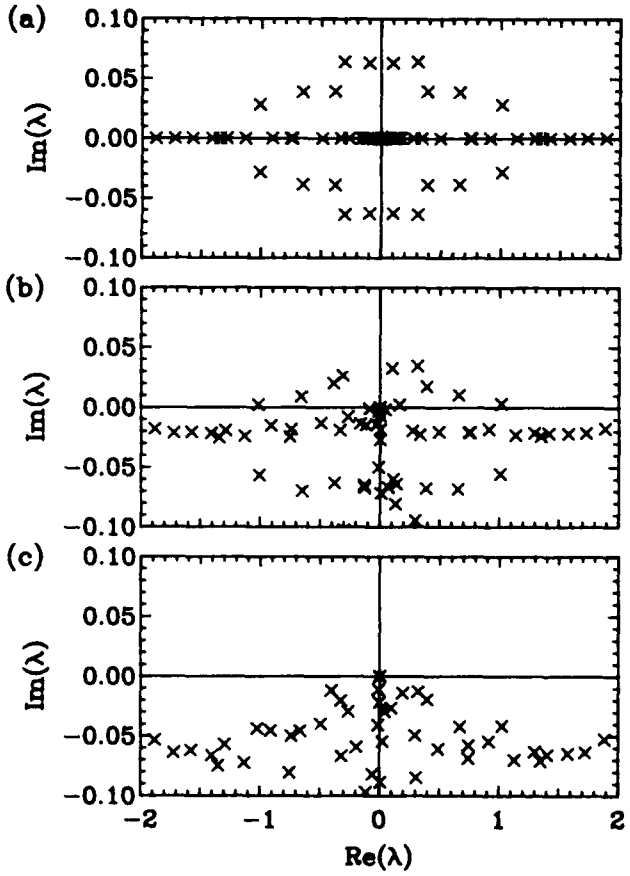


FIG. 6. Eigenvalue spectra for open systems using the boundary conditions of Eq. (3.4). If the boundary conditions are changed so as to open the system, nonzero imaginary parts are generated, as in (a). Because the boundary conditions are time reversible, these imaginary parts occur in conjugate pairs. If a damping term is added as in (b), most, but not all, imaginary parts are negative. The few eigenvalues with positive imaginary parts are sufficient to render the model unstable. Stability can be achieved by increasing the damping rate, leading to the spectrum (c).

IV. IRREVERSIBLE OPEN-SYSTEM MODEL

To provide a physical motivation for the ideas that openness necessarily involves time irreversibility, let us consider another example system drawn from electronic technology, the vacuum thermionic device ("vacuum tube" or "valve") (Langmuir and Compton, 1931; Eastman, 1949). These devices were made by introducing two or more metallic electrodes into a vacuum through which electrons could be transported without dissipation. When a voltage was applied between anode and cathode (and the cathode heated to thermally excite electrons into the vacuum), a nonequilibrium steady state would be established with a nonzero current flowing. Such a nonequilibrium steady state cannot be established in a reversible (or Hamiltonian) system. Consider what would happen if a population of electrons were introduced into some sort of trapping potential in ultrahigh vacuum. The system would effectively be closed, and the motion of the electrons would consist of periodic (thus, reversible) orbits. Of course what happened in the case of the thermionic vacuum tube is that electrons were accelerated by the electrostatic field until they impacted the anode, where they lost their kinetic energy to collisions with the electrons in the metal. Their energy was thus dissipated as heat. However, we can infer a much broader principle from this device: Making contact to a system in such a way as to permit particles to enter and leave (opening the system) in itself introduces irreversibility into the behavior of the system, so long as the contacts have a sufficient number of degrees of freedom and enough indistinguishable particles to behave as reservoirs.

Now, if the openness of the system is to be modeled by boundary conditions applied to the system, these boundary conditions must themselves be time irreversible. A physically appealing way to achieve such irreversibility is

to distinguish between particles moving into the system and those moving out of the system. It is then reasonable to expect that the distribution of particles flowing into the system depends only upon the properties of the reservoirs to which the system is connected, and that the distribution of particles flowing out of the system depends only upon the state of the system. The behavior of the reservoirs is thus analogous to that of an optical blackbody. This picture leads to a fully acceptable model of an open system.

A. Continuum formulation

To implement boundary conditions that distinguish between particles flowing into and those flowing out of a system, we must reexpress the Liouville equation (2.3) in terms of the classical phase space (q,p), where q in this case corresponds to the position x and p is the momentum. This is naturally done by the Wigner-Weyl transformation, which transforms the density operator $\rho(x,x')$ into the Wigner distribution function $f(q,p)$ (Wigner, 1932; Heller, 1976; Berry, 1977; Carruthers and Zachariasen, 1983). For the present purposes, the Wigner-Weyl transformation consists of a change of independent coordinates to the diagonal and cross-diagonal coordinates[1]:

$$q=\tfrac{1}{2}(x+x'),\quad r=x-x' , \tag{4.1}$$

followed by a Fourier transformation with respect to r. The variables x and x' may be expressed in terms of q and r by

$$x=q+\tfrac{1}{2}r,\quad x'=q-\tfrac{1}{2}r . \tag{4.2}$$

Thus the Wigner distribution can be expressed as

$$f(q,p)=\int_{-\infty}^{\infty} dr\,\rho(q+\tfrac{1}{2}r,q-\tfrac{1}{2}r)e^{-ipr/\hbar} . \tag{4.3}$$

The Liouville equation becomes

$$\frac{\partial f}{\partial t}=-\frac{p}{m}\frac{\partial f}{\partial q}-\frac{1}{\hbar}\int_{-\infty}^{\infty}\frac{dp'}{2\pi\hbar}V(q,p-p')f(q,p') , \tag{4.4}$$

where the kernel of the potential operator is given by

$$V(q,p)=2\int_{0}^{\infty} dr\,\sin(pr/\hbar)[v(q+\tfrac{1}{2}r)-v(q-\tfrac{1}{2}r)] . \tag{4.5}$$

These expressions are derived under the assumption that the domain is unbounded.

Let us consider the interpretation of the terms of the Liouville equation (4.4). The first term on the right-hand side is derived from the kinetic-energy operator and is of the form known as a drift, streaming, or advection term (in various nomenclatures). This term is exactly the same as the corresponding term of the classical Liouville equation with force F:

$$\frac{\partial f_{\rm cl}}{\partial t}=-\frac{p}{m}\frac{\partial f_{\rm cl}}{\partial q}-F\frac{\partial f_{\rm cl}}{\partial p} . \tag{4.6}$$

The correspondence between the classical and quantum drift terms will be exploited in defining the open-system boundary conditions.

Quantum-interference effects enter the Wigner-Weyl representation via the nonlocal potential term of Eq. (4.4). The kernel of this operator, $V(q,p-p')$, in effect redistributes the Wigner function among different p's at each position q. The extent to which it does so depends upon the potential at positions remote from q [Eq. (4.5)]. This is the way that interference between alternative paths is incorporated into the equation. Thus a rough intuitive image of the action of $V(q,p-p')$ is that it represents particles that have scattered off the potential at some point $q\pm\frac{1}{2}r$ and, upon returning, interfere with the particles propagating over other paths. This image will be invoked to interpret the effects of cutting off the integral in Eq. (4.5) at some finite value, which is required in practical computations.

Let us now consider a model in which the domain is bounded by $q=0$ and $q=l$. To address the question of boundary conditions, first note that in the Wigner-Weyl representation, the Liouville equation (4.4) is of first order with respect to q and does not contain derivatives with respect to p. The characteristics of the derivative term are lines of constant p, and we must supply one and only one boundary value at some point on each characteristic, because the equation is of first order on q. The kinds of boundary conditions that are appropriate are illustrated in Fig. 7. To implement the picture described above, that the particles entering the device depend only upon the state of the reservoirs and that the particles leaving the device depend only upon the state of the device, we should apply the boundary conditions illustrated in Fig. 7(c). That is, we set

$$\begin{aligned} f(0,p)|_{p>0}&=f^{(\rm left)}_{\rm boundary}(p) ,\\ f(l,p)|_{p<0}&=f^{(\rm right)}_{\rm boundary}(p) , \end{aligned} \tag{4.7}$$

where $f^{(\rm left)}_{\rm boundary}$ is the distribution function of the reservoir to the left of the system and $f^{(\rm right)}_{\rm boundary}$ is the distribution function of the reservoir to the right. These boundary conditions are not invariant under time reversal, because time reversal would change the problem of Fig. 7(c) into that of Fig. 7(d).

Conceptually, the boundary conditions (4.7) are identical to those employed in the conventional tunneling

[1]These are often referred to as "center of mass" and "relative" coordinates, respectively. I feel that this is a misleading terminology, because it gives the incorrect impression that one is dealing with a two-body problem. We shall see below that the significance of these coordinates follows from their relationship to the superoperators $\Upsilon_{+,-}$ and $\Upsilon_{-,+}$ [Eqs. (2.9), (2.10)] generated by the position operator.

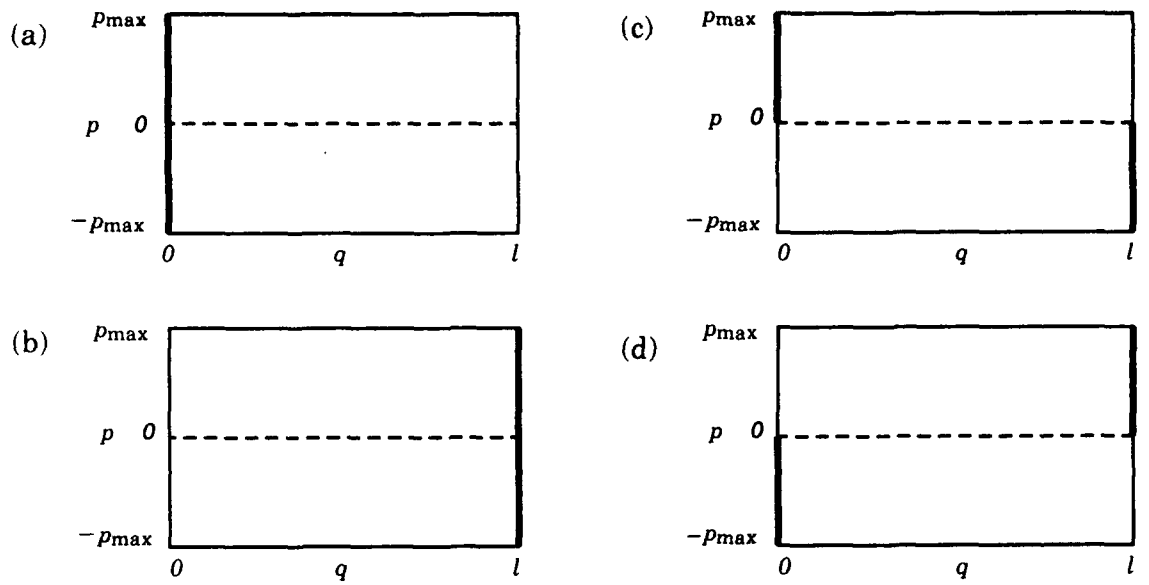


FIG. 7. Possible boundary conditions for the Liouville equation (4.4) in phase space. The points at which the boundary values are specified (indicated by a heavy line) can be at $q=0$ as in (a), at $q=l$ as in (b), or divided between the two boundaries, depending upon the sign of p, as shown in (c) and (d). The boundary conditions (c) are, in fact, the appropriate ones for an open system.

theory (see Appendices A and D), in the Landauer approach (Landauer, 1957, 1970; Büttiker *et al.*, 1985, Stone and Szafer, 1988), and in solutions of the Boltzmann equation for nonuniform systems (see Appendix C and Duderstadt and Martin, 1979). However, some care must be taken in this identification. It is true that the variable p goes over into the classical momentum appearing in the Boltzmann equation, by the correspondence principle. However, it is *not* true that p is the same quantity as the operator $p_x=(\hbar/i)\partial/\partial x$ or its eigenvalue. In particular, as will be discussed in Sec. VI.A, the traveling-wave boundary conditions actually depend upon the energy of the state, rather than p. Thus the boundary conditions (4.7) are conceptually identical to, but mathematically different from, those employed in the tunneling and Landauer approaches.

Let us call the Liouville superoperator which results from the boundary conditions (4.7) $\mathcal{L}^{(\mathrm{oi})}$ (for open system, irreversible). For purpose of the present discussion, it will be separated into two terms:

$$\mathcal{L}^{(\mathrm{oi})}=i\hbar\mathcal{T}+i\hbar\mathcal{V}\ , \tag{4.8}$$

where $\mathcal{T}$ is the superoperator derived from the kinetic-energy term of the Hamiltonian,

$$\mathcal{T}f=-\frac{p}{m}\frac{\partial f}{\partial q}\ , \tag{4.9}$$

and where $\mathcal{V}$ is the superoperator derived from the potential term,

$$(\mathcal{V}f)(q,p)=-\frac{1}{h}\int\frac{dp'}{2\pi\hbar}V(q,p-p')f(q,p')\ . \tag{4.10}$$

Let us note in passing that $\mathcal{V}$ can be written in two other forms. One is Groenewold's expression (Groenewold, 1946):

$$\mathcal{V}f(q,p)=\left[\frac{1}{i\hbar}\right]\left[v\left[q+\frac{i\hbar}{2}\frac{\partial}{\partial p}\right]-v\left[q-\frac{i\hbar}{2}\frac{\partial}{\partial p}\right]\right]f(q,p)\ . \tag{4.11}$$

The other is the Wigner-Moyal expansion (Moyal, 1949):

$$\begin{aligned}\mathcal{V}f(q,p)&=-\frac{2}{\hbar}\sum_{n=0}^{\infty}(-1)^n\frac{(\hbar/2)^{2n+1}}{(2n+1)!}\frac{\partial^{2n+1}v(q)}{\partial q^{2n+1}}\times\frac{\partial^{2n+1}f(q,p)}{\partial p^{2n+1}}\\&=-\frac{2}{\hbar}\sin\left[\frac{\hbar}{2}\frac{\partial}{\partial q}\frac{\partial}{\partial p}\right]v(q)f(q,p)\ ,\end{aligned} \tag{4.12}$$

where in the last expression it is understood that $\partial/\partial q$ acts only upon $v(q)$. The utility of both of these expressions depends upon the existence of a rapidly converging series expansion for $v(q)$. Such an expansion is not available for the abrupt energy-barrier structures that originally motivated the present study, so the integral form of $\mathcal{V}$ (4.10) is preferred for practical computations.

B. Discrete model

To investigate the eigenvalue spectrum of the Liouville operator subject to the boundary conditions (4.7) we

again construct a small, discrete model. The position variable q will take the same set of discrete values that x did in the previous section: $\{q_j | q_j = j\Delta_q$ for $j = 1,2,\ldots,N_q\}$. The values of p are also restricted to a discrete, bounded set: $\{p_k | p_k = (\pi\hbar/\Delta_q)[(k-\frac{1}{2})/N_p - \frac{1}{2}]$ for $k = 1,2,\ldots,N_p\}$. The mesh spacing in the p direction is thus $\Delta_p = (\pi\hbar)/(N_p\Delta_q)$. The choice of discrete values for p follows from a desire to avoid the point $p=0$ and the need to satisfy a Fourier completeness relation, which will be discussed later. The discrete Wigner distribution is then related to the discrete density matrix of Sec. III.B by

$$f_{jk} = \sum_{j'=-N_q/2}^{N_q/2} \rho_{j+j',j-j'} e^{-2ip_k j'\Delta_q/\hbar} , \tag{4.13}$$

where j indexes position q, and k indexes momentum p.

The discrete version of the potential term is readily defined. Using Eq. (4.13), we find that the discrete potential kernel becomes

$$V_{jk} = \frac{2}{N_p} \sum_{j'=1}^{N_q/2} \sin\left[\frac{2k\Delta_p j'\Delta_q}{\hbar}\right](v_{j+j'} - v_{j-j'}) . \tag{4.14}$$

[Notice that Eq. (4.14) invokes values of v_j that are outside the domain $\{q_j | j = 1,\ldots,N_q\}$. This expresses the nonlocality of quantum phenomena and is one way in which the environment of an open system influences the system's behavior. The values that one assumes for v_j, where $j<0$ or $j>N_q$, depend upon the nature of the environment. If ideal reservoirs are assumed, then setting these values equal to the potential at the appropriate boundary appears to be an adequate procedure.] The elements of $\mathcal{V}$ are then

$$\mathcal{V}_{jk;j'k'} = -\delta_{jj'} V_{j,(k-k') \bmod N_p}/\hbar = -\delta_{jj'} V_{j;k,k'}/\hbar , \tag{4.15}$$

where the notation $V_{j;k,k'} = V_{j,(k-k') \bmod N_p}$ is introduced to shorten the expressions to be derived from the discrete Liouville equation. Note that the elements of $\mathcal{V}$ are real and that $\mathcal{V}_{jk;j'k'} = -\mathcal{V}_{j'k';jk}$ so $(i\hbar\mathcal{V})$ is an imaginary Hermitian superoperator.

The boundary conditions (4.7) affect the form of the drift term $\mathcal{T}$ because they determine the proper finite-difference form for the gradient. On a discrete mesh, a first derivative $(\partial f/\partial q)(q_j)$ can be approximated by either a left-hand difference,

$$\left.\frac{\partial f}{\partial q}\right|_{\text{left}}(q_j) = \frac{f(q_j) - f(q_{j-1})}{\Delta_q} , \tag{4.16}$$

or a right-hand difference,

$$\left.\frac{\partial f}{\partial q}\right|_{\text{right}}(q_j) = \frac{f(q_{j+1}) - f(q_j)}{\Delta_q} . \tag{4.17}$$

(There is also a centered-difference form, $[f(q_{j+1}) - f(q_{j-1})]/2\Delta_q$, which has poor stability properties when used to approximate a drift term.) The boundary conditions determine which of the above difference forms must be used simply because one or the other will not couple the boundary value into the domain. Again, let us imagine that the boundary conditions (4.7) are implemented by fixing the value of f on mesh points just outside the domain:

$$\begin{aligned} f_{0,k} &= f^{(\text{left})}_{\text{boundary}_k} \quad \text{for } p_k > 0 , \\ f_{N_q+1,k} &= f^{(\text{right})}_{\text{boundary}_k} \quad \text{for } p_k < 0 . \end{aligned} \tag{4.18}$$

This scheme is illustrated in Fig. 8. Consider $p_k > 0$. The boundary conditions are specified for q_0, and if this value is to be coupled into the domain, we must use the left-hand difference formula (4.16) for the gradient at q_1. Consistency then requires that we use the left-hand difference for all q_j (for $p_k > 0$). Similarly, we must use the right-hand difference (4.17) for $p_k < 0$. In the context of hydrodynamic calculations such a difference scheme is called an "upwind" or "upstream" difference and is known to enormously enhance the stability of a computation (Roache, 1976, pp. 4–5). It has also been used in neutron transport calculations at the kinetic (phase space) level (Duderstadt and Martin, 1979). The elements of $\mathcal{T}$ are thus

$$\mathcal{T}_{jk;j'k'} = -\frac{p_k}{m\Delta_q}\delta_{kk'} \times \begin{cases} \delta_{j+1,j'} - \delta_{j,j'} & \text{for } p_k > 0 \\ \delta_{j,j'} - \delta_{j-1,j'} & \text{for } p_k < 0 . \end{cases} \tag{4.19}$$

The terms $\mathcal{T}_{1,k;0,k}$ and $\mathcal{T}_{N_q,k;N_q+1,k}$ couple to the fixed boundary values of f and are thus the coefficients of in-

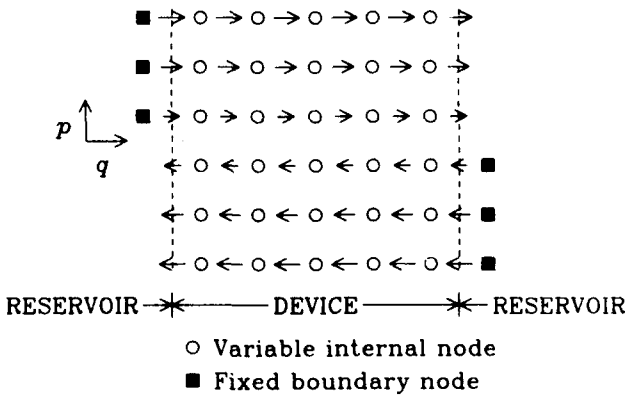


FIG. 8. Discretization scheme for the kinetic-energy superoperator (drift term) $\mathcal{T}$ in the Wigner representation. The flow of probability between mesh points is indicated by the arrows, which also define the sense of the finite-difference approximation for the gradient. A flow toward the right requires a left-hand difference and vice versa. This is the "upwind" difference scheme and is uniquely determined by the form of the boundary conditions (4.7).

homogeneous terms and are not strictly elements of $\mathcal{T}$. (In particular, these terms are not included in the eigenvalue calculation because eigenvalues are properties of homogeneous linear operators.) It is convenient to group these terms into a boundary contribution b_{jk}:

$$b_{1k} = \frac{p_k}{m\Delta_q} f^{(\text{left})}_{\text{boundary}_k} \quad \text{for } p_k > 0 ,$$
$$b_{N_q k} = -\frac{p_k}{m\Delta_q} f^{(\text{right})}_{\text{boundary}_k} \quad \text{for } p_k < 0 . \tag{4.20}$$

The discrete form of the Liouville equation then becomes

$$\frac{\partial f_{jk}}{\partial t} = \frac{1}{i\hbar} \sum_{j',k'} \mathcal{L}^{(\text{oi})}_{jk;j'k'} f_{j'k'} + b_{jk} , \tag{4.21}$$

with the inhomogeneous terms explicitly displayed. Expanding the definitions of the operators, the Liouville equation can be written as

$$\frac{\partial f_{j,k}}{\partial t} = -\frac{p_k}{m\Delta_q} \times \begin{cases} f_{j+1,k} - f_{j,k} & \text{for } p_k < 0 \\ f_{j,k} - f_{j-1,k} & \text{for } p_k > 0 \end{cases} - \frac{1}{\hbar} \sum_{k'} V_{j;k,k'} f_{j,k'} . \tag{4.22}$$

This provides a more convenient starting point for many of the manipulations that will be described below.

The eigenvalue spectrum for $\mathcal{L}^{(\text{oi})}$ constructed from Eqs. (4.8), (4.15), and (4.19) is shown in Fig. 9. The potential of Fig. 4 was used, with $N_q = 8$ and $N_p = 8$. All the eigenvalues of $\mathcal{L}^{(\text{oi})}$ have negative imaginary parts.

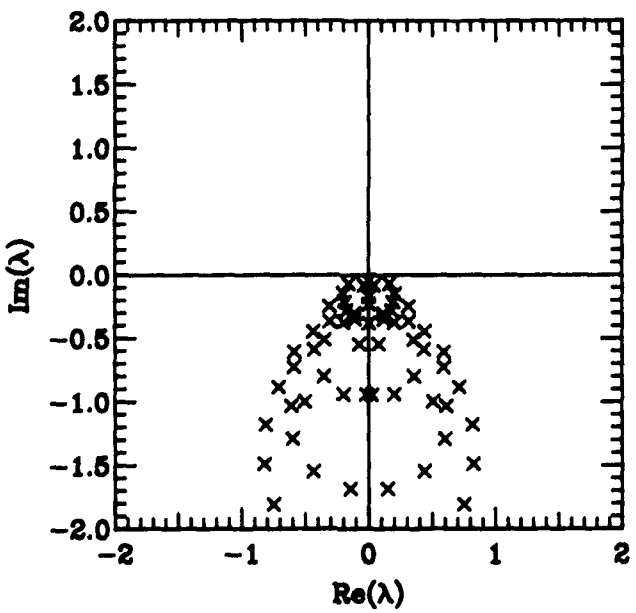


FIG. 9. Eigenvalue spectrum for a model open system with irreversible boundary conditions. All eigenvalues have negative imaginary parts, verifying that the model is stable, despite the fact that no damping is yet included.

(Note in particular that there is no eigenvalue equal to zero, and thus $\mathcal{L}^{(\text{oi})}$ is nonsingular). Because the eigenvalues have negative imaginary parts, the time dependence of f contains only decaying exponentials, so the model is stable. The stability of this model follows from the boundary conditions (4.7) and does not depend upon discretization (Frensley, 1986). To demonstrate this, let us consider the expectation value of $(\mathcal{L}^{(\text{oi})}/i\hbar)$ with respect to an arbitrary distribution f: $\langle f \| (\mathcal{L}^{(\text{oi})}/i\hbar) \| f \rangle$. If we demonstrate that this is nonpositive for any f, we will have shown that no eigenvalue of $(\mathcal{L}^{(\text{oi})}/i\hbar)$ has a positive real part, because the operator itself is purely real. In the Wigner-Weyl representation the operator inner product (2.4) becomes simply (Wigner, 1971; Hillery, O'Connell, Scully, and Wigner, 1984)

$$\langle f \| g \rangle = \frac{1}{2\pi\hbar} \int dq \int dp \, f(q,p) g(q,p) . \tag{4.23}$$

The expectation value can be rewritten

$$\langle f \| (\mathcal{L}^{(\text{oi})}/i\hbar) \| f \rangle = \langle f \| \mathcal{T} \| f \rangle + \langle f \| \mathcal{V} \| f \rangle = \langle f \| \mathcal{T} \| f \rangle , \tag{4.24}$$

because $\langle f \| \mathcal{V} \| f \rangle = 0$ from the antisymmetry of $\mathcal{V}$. For the mathematically homogeneous problem (source terms set to zero) the boundary conditions are $f(0,p) = 0$ for $p > 0$ and $f(l,p) = 0$ for $p < 0$. With this we can integrate the expectation value for $\mathcal{T}$ and simplify it to obtain

$$\langle f \| \mathcal{T} \| f \rangle = \frac{1}{4\pi\hbar m} \left| \int_{-\infty}^{\infty} p f^2(0,p) dp - \int_{-\infty}^{\infty} p f^2(l,p) dp \right| = \frac{1}{4\pi\hbar m} \left| \int_{-\infty}^{0} p f^2(0,p) dp - \int_{0}^{\infty} p f^2(l,p) dp \right| \le 0 . \tag{4.25}$$

Thus the stability of the solutions to the Liouville equation using $\mathcal{L}^{(\text{oi})}$ follows from the boundary conditions alone. The physical significance of this argument is that the particles in an open system will eventually escape and the density will approach zero if there is no inward current flow from the environment. However, if the potential has a local minimum within the system deep enough to create one or more bound states, any particles in those states will not escape. Their contributions to f will be zero at the boundaries, and this is the significance of the case in which Eq. (4.25) is equal to zero. Such states should correspond to eigenvalues of $\mathcal{L}^{(\text{oi})}$ that are equal to zero, although I have not observed such a situation in the models that I have examined. In an open system of finite extent and with potentials of finite depth, the tunneling tail of bound-state wave function will be nonzero at the system boundaries, perhaps leading to a finite rate of escape from that state within the present model.

Let us examine how this open-system model can be used. The methods of calculation are more readily visualized if we write Eq. (4.21) in a block-matrix notation:

$$\frac{\partial}{\partial t}\begin{vmatrix}[f]_1\\[f]_2\\\vdots\\[f]_{N_q}\end{vmatrix}=\begin{vmatrix}[\mathcal{T}+\mathcal{V}]_{11} & [\mathcal{T}]_{12} & & & \\ [\mathcal{T}]_{21} & [\mathcal{T}+\mathcal{V}]_{22} & [\mathcal{T}]_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & [\mathcal{T}]_{N_q,N_q-1} & [\mathcal{T}+\mathcal{V}]_{N_q,N_q}\end{vmatrix}\begin{vmatrix}[f]_1\\[f]_2\\\vdots\\[f]_{N_q}\end{vmatrix}+\begin{vmatrix}[b]_1\\0\\\vdots\\[b]_{N_q}\end{vmatrix}. \tag{4.26}$$

Here $[f]_j$ and $[b]_j$ represent column vectors, and $[\mathcal{T}]_{jj'}$ and $[\mathcal{V}]_{jj}$ represent matrices, whose internal indices range over the allowed values of k. The $[\mathcal{T}]$ are diagonal matrices, whereas the $[\mathcal{V}]$ are dense. The block-tridiagonal form of $\mathcal{L}^{(oi)}$ greatly reduces the computational labor required to solve the Liouville equation as compared to that required to work with superoperators of a more general form.

Now suppose that we wish to find the nonequilibrium steady state ($\partial f_{jk}/\partial t=0$). Can we simply move the $[b]_j$ column vector over to the other side of the equation and solve for the f_{jk}? The answer is yes, provided that the operator $\mathcal{L}^{(oi)}$ is nonsingular. If there are no bound states, all the eigenvalues of $\mathcal{L}^{(oi)}$ are nonzero (see Fig. 9), so $\mathcal{L}^{(oi)}$ is a nonsingular operator and its inverse exists. This steady-state solution for the Wigner function may be written

$$f^{(dc)}=-i\hbar\mathcal{L}^{(oi)\,-1}b\ , \tag{4.27}$$

where $f^{(dc)}$ refers to the "direct-current" case. Equation (4.26) is also used to solve time-dependent problems, as will be described in the following section.

Let us compare this approach to the most commonly studied problem in transport theory, transport in a spatially homogeneous system with a uniform driving field (as is done to evaluate transport coefficients such as mobilities) (Dresden, 1961; Conwell, 1967). This generates a mathematically homogeneous problem, and the solution corresponds to the null space of that superoperator which appears in the transport equation (Aubert, Vaissiere, and Nougier, 1984). Thus the superoperator must be singular and, if the transport equation is linear, the solution is not unique (the total density is not determined). What the present model demonstrates is that this formulation of transport through a spatially inhomogeneous system leads to a mathematically inhomogeneous problem, which is in many respects a good deal simpler than a similar homogeneous problem. For example, because $\mathcal{L}^{(oi)}$ is nonsingular, there is no problem of compatibility relations for the boundary conditions (Lanczos, 1961). Any choice of distribution function on the boundary will generate a unique steady-state solution. The same considerations apply to the evaluation of the transient response of an open system by integrating Eq. (4.4) with respect to t. The solution in unique and, as we have seen, stable.

These considerations clarify a point discussed by Kluksdahl *et al.* (1989), concerning the role of the initially assumed Wigner function in a calculation in which the steady state is found by simulating the time evolution. Kluksdahl *et al.* assert that the initial state must be quantum-mechanically correct. The only components of the initial state that remain through the time-evolution calculation are those lying in the null space of the Liouville operator. All other components will approach steady-state values that are independent of the initial condition. Thus, if there is no null space (the operator is nonsingular), the initial condition makes no difference whatsoever. A concern about the correctness of the initial state is warranted only if there are bound states within the system, and possibly in the continuum limit where the smallest eigenvalue approaches zero.

V. APPLICATION OF THE IRREVERSIBLE MODEL TO TUNNELING DIODES

To illustrate the application of this irreversible open-system model to a specific physical system, let us consider the semiconductor heterostructure resonant-tunneling diode (RTD; Chang, Esaki, and Tsu, 1974; Sollner *et al.*, 1983). The study of this device provided the original motivation for the present investigation. The RTD exploits the ability of modern heteroepitaxial technologies to grow extremely thin layers of chemically different semiconductors (such as gallium arsenide, GaAs, and aluminum arsenide, AlAs) on top of one another in a single crystal structure. To a surprising degree of accuracy, the effects of such a structure on the motion of free electrons (or holes) may be modeled by an effective potential that is related to the local energy-band gap and is thus a function of the local chemical composition (Dingle, Wiegmann, and Henry, 1974). Therefore a structure consisting of a layer of GaAs a few nanometers thick placed between layers of AlAs (or more commonly a solid solution $Al_xGa_{1-x}As$ with $x\approx 0.3$) forms a rectangular potential well of finite depth for electrons. The shift in energy due to size quantization of the states in the well is enormously enhanced by the low effective mass of electrons in GaAs (0.067 of the free-electron mass), so the same shift is obtained in quantum wells tens of atomic layers thick in GaAs as would be obtained in structures of atomic dimensions in free space.

The behavior of the resonant-tunneling diode is summarized in Fig. 10. The device consists of a quantum well bounded by barrier layers thin enough to permit tunneling. Outside the barrier layers are thick layers of lower effective potential, which are doped so as to have a significant density of free electrons and to which electrical contact is made. The confined states in the quantum well thus become resonances in this structure, and electrons may readily tunnel through these resonances only if they have the correct energy. The energy of the reso-

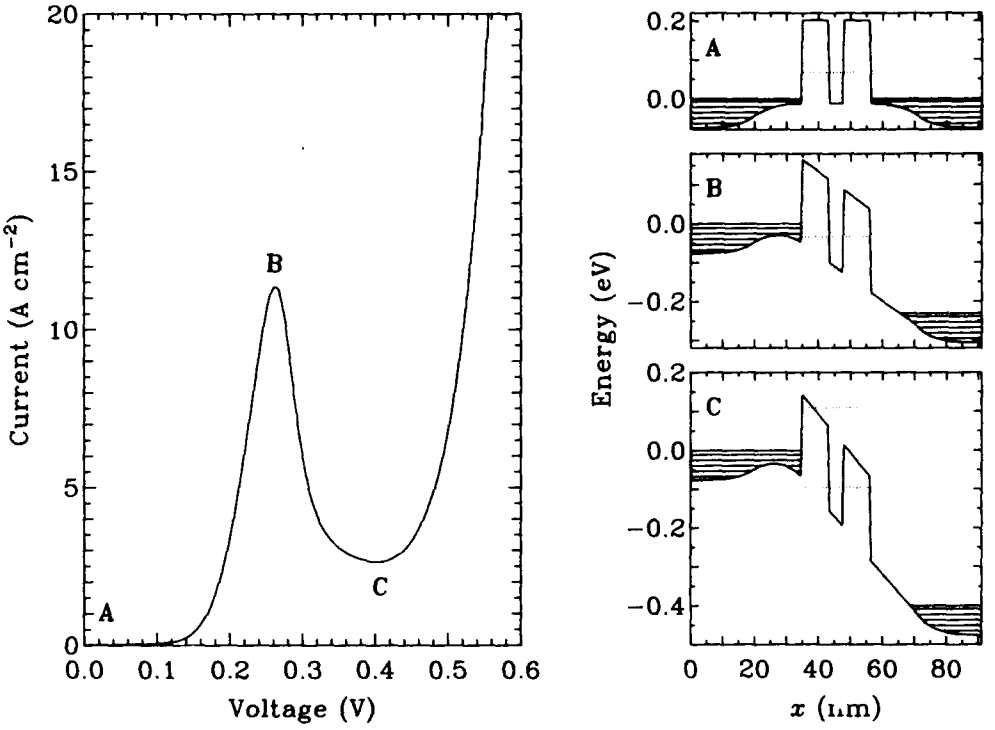


FIG. 10. Summary of the properties of the quantum-well-resonant-tunneling diode. The $J(V)$ curve of an experimental device (Reed *et al.*, 1989) at a temperature of 77 K is shown to the left. The diagrams to the right show the conduction-band profile of the device at different bias voltages corresponding to the noted points on the $J(V)$ curve. The shaded regions show the occupied electron states. In equilibrium (A) the current is zero. As a bias voltage is applied, the resonant level (dotted line) is pulled down in energy so that it lines up with the occupied electron states, permitting resonant tunneling (B). As the voltage is increased, the resonant level eventually passes below the lowest occupied state in the cathode (left-hand electrode), and the resonant-tunneling current ceases (C). The current subsequently increases as conduction through higher-energy states becomes possible. The rise in the conduction-band potential near the quantum well apparent in (A) is the result of a nonuniform distribution of impurity ions, which is a part of the design of the device.

nances varies with externally applied electrostatic potential. In particular, at a sufficiently large bias voltage the resonance is pulled below the lowest occupied state in the cathode layer and the resonant-tunneling current ceases. This leads to a decreasing current with increasing voltage ("negative differential resistance"), which is an unambiguous indication of resonant tunneling in this structure.

Over the past few years a great deal of work concerning the resonant-tunneling diode, both theoretical and experimental, has been published. Most of the theoretical treatments are expressed in terms of the transmission probabilities associated with pure quantum states. Due to the volume of this work, no attempt will be made to review it comprehensively here, but we shall instead concentrate upon the kinetic models.

A. Steady-state (dc) behavior

The steady-state behavior of the RTD has been evaluated using the Wigner function in an open-system model by several groups (Frensley, 1986, 1987; Kluksdahl *et al.*, 1988, 1989; Mains and Haddad, 1988b; Jensen and Buot, 1989a). To obtain the results described here, the steady-state Wigner function was evaluated using Eq. (4.27) repeatedly for a set of potentials representing different applied bias voltages. (The assumed structure consisted of a 4.5 nm GaAs quantum well bounded by 2.8 nm $Al_{0.3}Ga_{0.7}As$ barrier layers. The contact layers were assumed to be doped so as to produce a free-electron density of 2×10^{17} cm^{-3}, and the temperature was taken to be 300 K.) The boundary distribution was taken to be

$$f_{\text{boundary}}(p_k)=(m^*/\pi\hbar^2\beta)\times\ln[1+e^{-\beta(p_k^2/2m^*+v-\mu)}], \qquad (5.1)$$

to include the integration over transverse momenta. [Here $v-\mu$ is evaluated at each boundary using the charge-neutrality condition (A5).] The current density was evaluated from $f^{(\text{dc})}$, and the resulting $J(V)$ curve is plotted in Fig. 11. Also shown for comparison is the result of a more conventional tunneling theory calculation, such as that described in Appendix A. [More specifically, it is the current density that would be obtained by taking the expectation value of the current

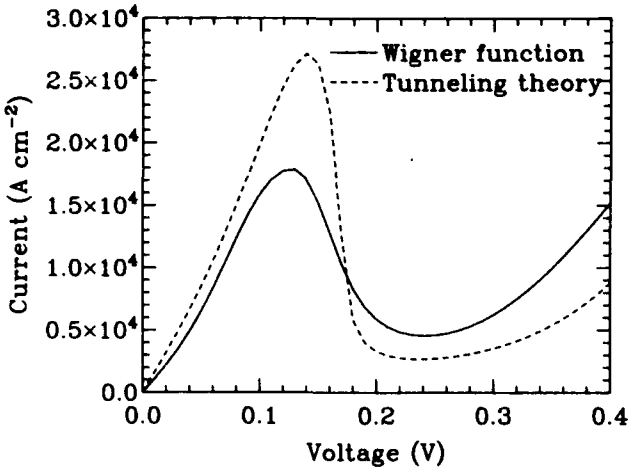


FIG. 11. Current density as a function of voltage for a model resonant-tunneling diode. The result of the time-irreversible kinetic (Wigner function) model is shown by the solid line, and a more conventional tunneling calculation is shown by the dashed line. While they differ in detail, the calculations agree as to the qualitative behavior of the far-from-equilibrium steady state and predict tunneling currents of the same order of magnitude.

operator with respect to the density operator (A6).] The two calculations agree on the qualitative shape of the $J(V)$ curve and on the voltages at which the peak and valley occur. There is a disagreement of some tens of percent on the magnitude of the peak and valley current.

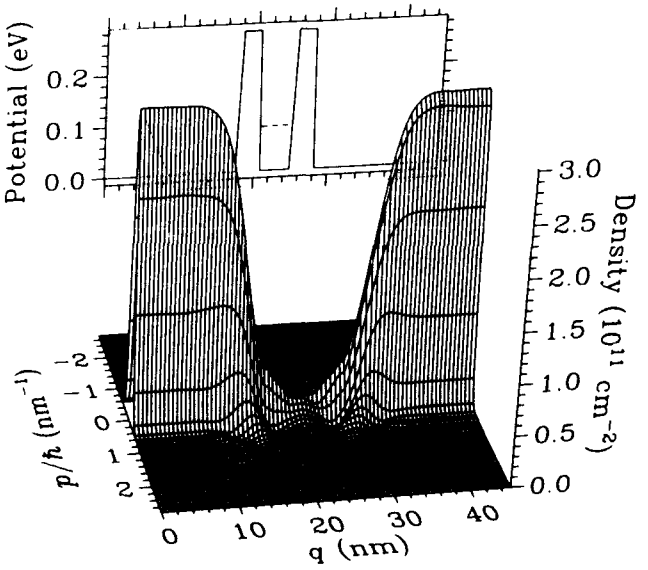


FIG. 12. Wigner distribution function for the resonant-tunneling diode at zero bias voltage (thermal equilibrium). In the electrode (flat-potential) regions the distribution is approximately Maxwellian (as a function of p). The density is reduced in the vicinity of the quantum well due to size-quantization effects. The very small ripples perceptible at larger p are due to standing waves near the energy barriers.

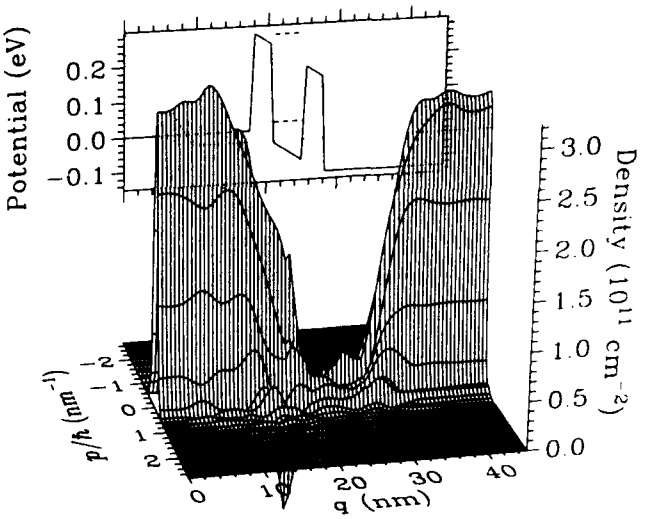


FIG. 13. Wigner distribution function for 0.13 V bias, at the peak of the $J(V)$ curve. The complex standing-wave patterns and prominent negative peak indicate that strong quantum-interference effects are present.

One can cite at least two possible sources of this disagreement. The more obvious one is that the Wigner-function calculation necessarily introduces a limited coherence length because in the discrete approximation the integral defining the nonlocal potential (4.5) must be cut off at a finite value as in Eq. (4.14). The tunneling theory is based upon solutions of Schrödinger's equation, which necessarily assumes an infinite coherence length. A second, and probably more fundamental, explanation for the disagreement is that the tunneling and kinetic theories are simply not equivalent (the kinetic theory being Markovian while the tunneling theory is not). The notion that these theories can be viewed as different approximations to a more general many-body theory is ex-

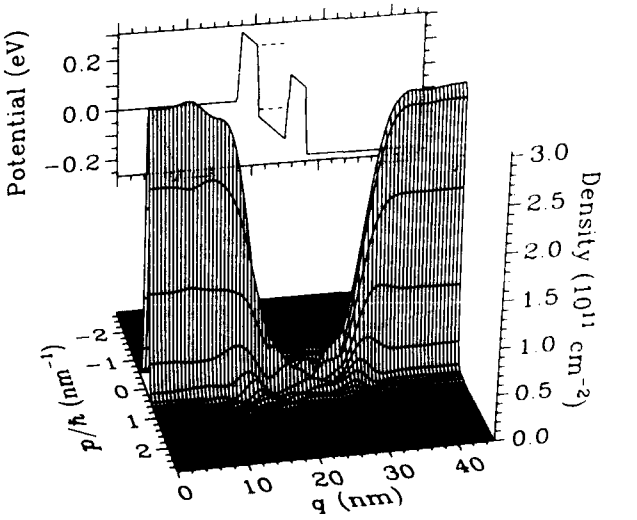


FIG. 14. Wigner distribution function for 0.24 V bias, corresponding to the bottom of the valley in the $J(V)$ curve. This case is quite similar to the equilibrium case of Fig. 12.

amined in Sec. VI.C.

The Wigner distribution functions that underlie the $J(V)$ curve of Fig. 11 are illustrated in Figs. 12–14. The equilibrium (zero-bias) case is shown in Fig. 12. The large electron density in the electrode regions, and much smaller density in the vicinity of the quantum well, is evident. Figure 13 shows the Wigner function for a bias voltage of 0.13 V, which corresponds to the peak of the resonant-tunneling current. The negative peak indicates that strong quantum-interference effects are present. In contrast, the Wigner function for 0.24 V, at the minimum valley current, is quite similar to the equilibrium case.

B. Large-signal transient response

As discussed in Sec. II.D, a principal reason for adopting a kinetic-level model is the desire to evaluate the time evolution of an irreversible system. Again, this has been demonstrated using open-system Wigner-function models (Ravaioli *et al*, 1985; Frensley, 1986, 1987a; Kluksdahl *et al.*, 1988). As an example, let us consider abruptly changing the bias voltage on the model RTD. Then the Wigner function f will initially equal the steady-state value at the first bias voltage. After the voltage is changed, f will evolve and approach the steady-state value at the new bias voltage. This time evolution may be evaluated by integrating Eq. (4.22), now regarding the potential as a time-dependent quantity. The integration with respect to t is readily done by discretizing t in units Δ_t. For purely numerical considerations of stability (see Frensley, 1987a), an effective way to implement the time integration is using the "fully implicit" or "backward Euler" approach, which involves repeatedly solving

$$[f(t+\Delta_t)-f(t)]/\Delta_t=(\mathcal{L}^{(\mathrm{oi})}/i\hbar)f(t+\Delta_t)+b\ , \qquad (5.2)$$

to advance the solution for $f(t)$ forward in time. This is equivalent to expanding the exponential of the Liouville operator in a product expansion,

$$\exp(-i\mathcal{L}^{(\mathrm{oi})}t/\hbar)\approx(1+i\mathcal{L}^{(\mathrm{oi})}t/n\hbar)^{-n}\ . \qquad (5.3)$$

Note that, because $\mathcal{L}^{(\mathrm{oi})}$ is not Hermitian, $\exp(-i\mathcal{L}^{(\mathrm{oi})}t)$ is not unitary. It is thus not necessary to use the unitarity-preserving Cayley (or Crank-Nicholson) form,

$$e^{-iHt}\approx\left|\frac{1-iHt/2n}{1+iHt/2n}\right|^n\ ,$$

which is preferred for the integration of Schrödinger's equation. The fully implicit scheme is a bit simpler to implement (and to explain) than the Cayley scheme, but the latter will generally be more accurate (see Jensen and Buot, 1989a) and probably should be preferred.

The transient-response calculation was carried out (using the fully implicit scheme) for the particularly interesting case in which the RTD is suddenly switched across the negative-resistance region. The spatially averaged current density (which would equal the current induced in the external circuit, apart from parasitic effects) is

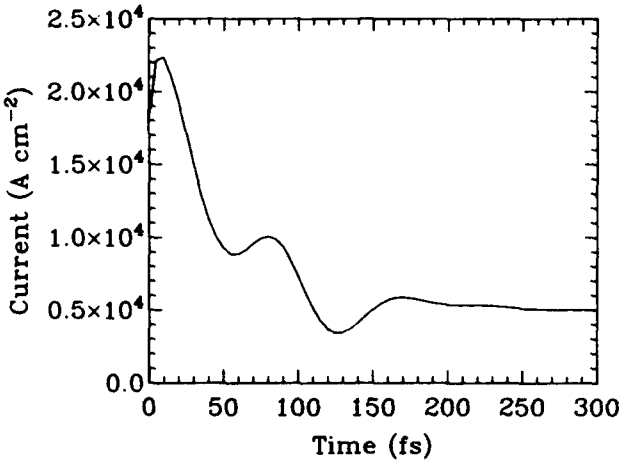


FIG. 15. Results of a calculation of the transient response of the resonant-tunneling diode. For $t<0$ the device was in steady state at $V=0.13$ V, the peak of the $J(V)$ curve of Fig. 11. At $t=0$ the voltage was switched to $V=0.24$ V, the bottom of the valley. The conduction current density averaged over the device (which equals the current induced in the external circuit) is plotted as a function of t. The current initially increases and then declines with some superimposed oscillations toward the new steady state. Parasitic effects are neglected.

plotted in Fig. 15. The current initially rises in response to the increased field and then decreases toward its steady-state value with some superimposed oscillations. More insight can be gained into the transient process by plotting the current density as a function of both time and position within the device as in Fig. 16. There is an initial peak within the quantum well, which reflects the shifting electron distribution in response to the increased field. The current density in the downstream part of the

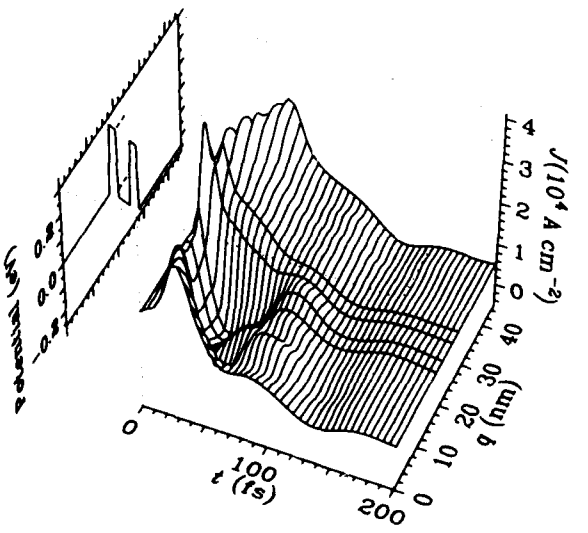


FIG. 16. The same transient-response calculation as that shown in Fig. 15, but here the current density is shown as a function of position q within the device.

device then declines fairly monotonically, presumably reflecting a simple single-barrier tunneling process which empties the quantum well. On the upstream side of the structure the current transient is much more oscillatory. The reason for this is presumably the change in reflection coefficient caused by the shift in the potential and the resulting transient changes in the standing-wave patterns in this region. The most significant result of the calculation, however, is the demonstration of a stable approach to steady state.

C. Small-signal ac response

Another aspect of the behavior of electronic devices which is of much interest to circuit designers is the small-signal ac response of the device. This is the response of the device to a small sinusoidal voltage imposed upon a generally much larger dc bias voltage. That is, one seeks to evaluate the effect of a small perturbation on a far-from-equilibrium steady state. This is a rather different problem from that treated by the linear-response theory of statistical physics (Kubo, 1957), which seeks to evaluate the effect of small perturbations on an equilibrium state. A perturbation expansion of the present kinetic theory may be readily obtained to evaluate the small-signal ac response of our model RTD (Frensley, 1987b, 1988a; Mains and Haddad, 1988b). Let us assume that the potential of the system varies as

$$v(x,t)=v_0(x)+\tfrac{1}{2}\lambda[v_\omega(x)e^{i\omega t}+\text{c.c.}]\ , \tag{5.4}$$

where c.c. denotes the complex conjugate, $v_0(x)$ is the dc potential including the heterostructure and the large bias voltage, $v_\omega(x)$ is the potential due to the small ac voltage, and λ is a perturbation parameter introduced solely to keep track of the order of the perturbation (and is ultimately set equal to unity). We should expect that the current induced in the external circuit can be expanded as

$$\begin{aligned} I(t)=&I_0(V_0)+\tfrac{1}{2}\lambda[y(\omega)V_\omega e^{i\omega t}+\text{c.c.}] \\ &+\tfrac{1}{2}\lambda^2 a_{\text{rect}}(\omega)V_\omega^2+\tfrac{1}{4}\lambda^2[a_{2\omega}(\omega)V_\omega^2 e^{2i\omega t}+\text{c.c.}] \\ &+\cdots\ , \end{aligned} \tag{5.5}$$

where $V_0=[v_0(l)-v_0(0)]/e$ and $V_\omega=[v_\omega(l)-v_\omega(0)]/e$ are the total voltages applied, e being the charge of the electron. The coefficients of Eq. (5.5) describe different aspects of the ac response: y is the linear admittance, the amount of rectification of the sinusoidal wave form is given by a_{rect}, and the amount of second-harmonic generation is given by $a_{2\omega}$. Note that at $\omega=0$ these coefficients are just the derivatives of the $I(V)$ curve: $y(0)=dI/dV$ and $a_{\text{rect}}(0)=a_{2\omega}(0)=d^2I/dV^2$. The coefficients of Eq. (5.5) at an arbitrary frequency may be obtained from the corresponding components of the Wigner function. To do this we write the Liouville operator as

$$\mathcal{L}^{(\text{oi})}(t)=\mathcal{L}_0^{(\text{oi})}+\tfrac{1}{2}i\hbar\lambda(\mathcal{V}_\omega e^{i\omega t}+\text{c.c.})\ . \tag{5.6}$$

The Wigner function can be expanded (to second order in λ) as

$$\begin{aligned} f(t)=&f^{(\text{dc})}+\tfrac{1}{2}\lambda(f^{(\omega)}e^{i\omega t}+\text{c.c.})+\lambda^2 f^{(\text{rect})} \\ &+\tfrac{1}{2}\lambda^2(f^{(2\omega)}e^{2i\omega t}+\text{c.c.})+\cdots\ . \end{aligned} \tag{5.7}$$

Inserting Eqs. (5.6) and (5.7) into the Liouville equation and collecting terms of equal frequency and order in λ leads to these equations:

$$f^{(\omega)}=-\frac{i\hbar}{\mathcal{L}_0^{(\text{oi})}+\hbar\omega}\mathcal{V}_\omega f^{(\text{dc})}\ , \tag{5.8}$$

$$f^{(\text{rect})}=-\frac{i\hbar}{2\mathcal{L}_0^{(\text{oi})}}\text{Re}(\mathcal{V}_\omega f^{(\omega)*})\ , \tag{5.9}$$

$$f^{(2\omega)}=-\frac{1}{2}\frac{i\hbar}{\mathcal{L}^{(\text{oi})}+2\hbar\omega}\mathcal{V}_\omega f^{(\omega)}\ , \tag{5.10}$$

where $f^{(\text{dc})}$ is obtained from Eq. (4.27). (The denominators of this perturbation series look a bit unfamiliar, with expressions of the form $\mathcal{L}+i\hbar\omega$ rather than $\mathcal{L}-i\hbar\omega$. The reason for this is that we have mixed the quantum-mechanical convention for the time dependence, $e^{-i\mathcal{L}t/\hbar}$, with the convention used in electronics, $e^{i\omega t}$. While a consistently quantum-mechanical notation would produce more conventional expressions, it would also produce a great deal of confusion when we examine the imaginary parts of the response to determine whether they resemble capacitances or inductances.) The superoperator resolvent expressions in Eqs. (5.8)–(5.10) are readily evaluated with the same algorithms used to solve the steady-state and transient problems.

Evaluating the expectation value of the current density J for any of the terms of $f(t)$ gives the conduction current as a function of position q:

$$\langle Jf^{(i)}\rangle(q)=\int_{-\infty}^{\infty}\frac{dp}{2\pi\hbar}\frac{p}{m}f^{(i)}(q,p)\ . \tag{5.11}$$

The current induced in the external circuit by this conduction current within the device is obtained by invoking the Shockley-Ramo theorem (Shockley, 1938; Ramo, 1939). We shall approximate the properties of the doped contacting layers as ideally metallic conductors bounded by interfaces to the higher-potential barrier layers at q_l and q_r. The Shockley-Ramo theorem then takes the form

$$I[f^{(i)}]=\frac{A}{q_r-q_l}\int_{q_l}^{q_r}dq\,\langle Jf^{(i)}\rangle(q)\ , \tag{5.12}$$

where A is the area of the device. The coefficients of the expansion of $I(t)$ (5.5) are thus given by

$$I_0(V_0)=I[f^{(\text{dc})}]\ , \tag{5.13}$$

$$y(\omega)=I[f^{(\omega)}]/V_\omega\ , \tag{5.14}$$

$$a_{\text{rect}}(\omega)=\tfrac{1}{2}I[f^{(\text{rect})}]/V_\omega^2\ , \tag{5.15}$$

$$a_{2\omega}(\omega)=\tfrac{1}{2}I[f^{(2\omega)}]/V_\omega^2\ . \tag{5.16}$$

It should be emphasized that these expressions represent only the conduction current component; the displacement current must be added to them to obtain a complete description of the behavior of the device.

The linear admittance y of the present RTD model was evaluated using Eqs. (5.8) and (5.14) at a bias of 0.17 V (in the middle of the negative-resistance region), as a function of frequency over the GHz and THz regions. The results are plotted in Fig. 17. The conductance Re(y) is negative at lower frequencies, as we would expect from the dc results. This negative conductance "rolls off" and becomes positive at about 6 THz, which is therefore the maximum frequency of oscillation of the intrinsic device (not including parasitic effects). The susceptance Im(y) is positive and proportional to ω at lower frequencies, which is the behavior of a capacitance. Recall, however, that the displacement current that flows through the geometrical device capacitance is not included in this calculation. The result that $\mathrm{Im}(y)>0$ is somewhat surprising, since the most obvious reactive effect in electron transport at high frequencies is the electron inertia, which leads to Im[$y(\omega)$] resembling that of an inductor with Im(y) negative (Champlin, Armstrong, and Gunderson, 1964). The initial calculations of the admittance by the present author (Frensley, 1987b, 1988a) gave negative Im(y) due to a programming error, and the electron-inertia explanation was proposed in those papers. During the preparation of the present work the error was discovered, and correcting it brings the results into agreement with those obtained by Mains and Haddad (1988b), who obtained positive Im(y). Thus the electron inertia does not explain the behavior of Im[$y(\omega)$], and an alternative explanation must be sought. A key piece of evidence is provided by evaluating the admittance of structures with either one energy barrier or none, in addition to the double-barrier structure. These structures do indeed show negative (inductive) Im(y), presumably due to electron inertia. The capacitive Im(y) is thus uniquely associated with the double-barrier structure and therefore must reflect the confinement of electrons in the quantum well. The idea that electron storage in a quantum well could be represented as a capacitance was proposed by Luryi (1985), but he identified this capacitance with the geometrical capacitance of the device, through which the displacement current flows. The storage capacitance inferred from the present calculation is 1–2 orders of magnitude smaller than the geometrical capacitance.

The rectification and second-harmonic generation coefficients a_{rect} and $a_{2\omega}$ were evaluated using Eqs. (5.9), (5.10), (5.15), and (5.16) at a bias of 0.13 V (the top of the current peak). The moduli of these quantities are shown in Fig. 18. While $a_{2\omega}$ decreases at higher frequencies, a_{rect} shows a resonant enhancement over the frequency range of 1 to 8 THz. This is quite interesting, because a_{rect} was measured by Sollner *et al.* (1983) at a frequency of 2.5 THz. The experimental data show that for most bias voltages $|a_{\mathrm{rect}}(2.5\ \mathrm{THz})|$ exceeds the dc $|d^2I/dV^2|$, indicating that the magnitude of a_{rect} must increase in this frequency range. On the other hand, the rectification process in the RTD has been recently analyzed by Wingreen (1990), using a transmission-coefficient ap-

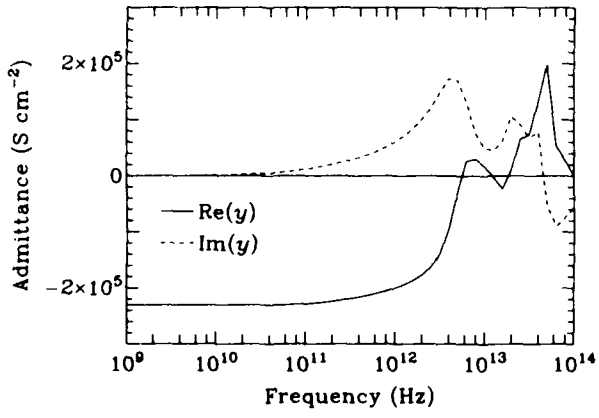


FIG. 17. Small-signal ac response of the resonant-tunneling diode for a dc bias of 0.17 V, which places the device in the middle of the negative-resistance region. The device conductance (the real part of the admittance, solid line) is negative at lower frequencies, with a value equal to that expected from the derivative of the dc $J(V)$ curve. The negative conductance decreases in magnitude and becomes positive at a few THz. The complex behavior at higher frequencies is an indication that optical transitions are becoming important. The susceptance (imaginary part of the admittance, dashed curve) has the same sign as a capacitance and is due to the effects of electron storage in the quantum well. These quantities reflect only the conduction current and do not include the displacement current through the parasitic capacitance of a real device. This displacement current would prevent observation of the higher-frequency effects in a realistic experimental situation.

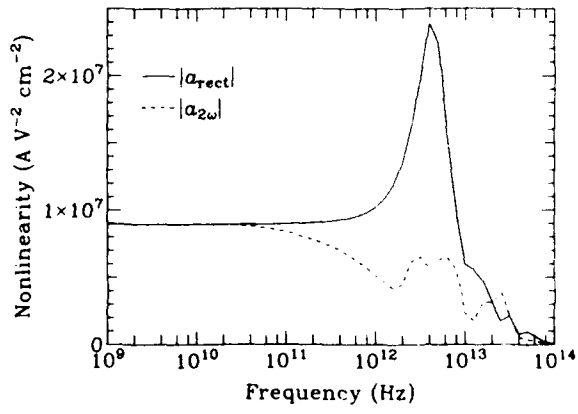


FIG. 18. Nonlinear response of the resonant-tunneling diode at a dc bias of 0.13 V, at the peak of the $J(V)$ curve. The rectification coefficient (solid line) shows a resonant enhancement near 6 THz.

proach. He found no evidence of enhancement, only a decrease in a_{rect} as the frequency is raised. One difference between Wingreen's calculation and that based upon Eq. (5.9) is that the former includes the effects of only one resonant level, whereas the latter includes all such levels. This suggests that the enhancement of a_{rect} might involve transitions between resonant levels, though the frequency of the transition between the lowest two levels in the present example is 60 THz, which argues against this notion. This illustrates one of the problems with a kinetic approach that incorporates all physical processes: Such an approach provides little guidance when one desires to identify that process which is the cause of some particular effect.

It is particularly interesting to look at $\langle Jf^{(\omega)} \rangle$ as a function of both frequency and position q. This is plotted in Fig. 19. At frequencies below a few THz the current is independent of position, as one would expect in an electron device. As the frequency increases above this value, the ac current density becomes strongly nonuniform, indicating that the response of the current to the applied potential is strongly nonlocal. A particularly prominent peak occurs in Re$[y(q)]$ at a frequency of 50 THz and centered within the quantum well. The positive value of the conductance in this peak indicates that the in-phase current density is locally large, so this part of the device is absorbing power from the ac electric field. The obvious explanation for this absorption is that the peak reflects quantum transitions between the two lowest resonances in the well. A transmission-coefficient calculation indicates that, for the present example, these states are separated in energy by 0.248 eV, for which the corresponding photon frequency is 60 THz. The small discrepancy in predicted frequencies is presumably attributable to the effect of the Markov assumption in the kinetic theory, as in the case of the $J(V)$ curves. Figure 19 is interesting because it gives us a view of the transition of a single system from the domain of electronics to that of optics.

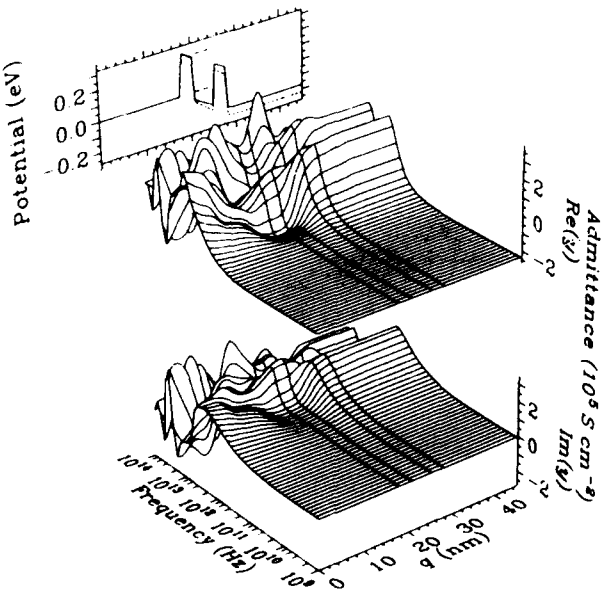


FIG. 19. Linear component of the ac current density (divided by the applied ac voltage and thus expressed as an admittance) as a function of frequency and position. At lower frequencies the current density is spatially uniform, but strong nonlocal effects develop as the frequency is increased. This is a characteristic of the transition from electronic to optical behavior. The prominent peak in Re(y) centered in the quantum well at 50 THz is due to quantum transitions between the two lowest resonant levels.

In addition to these effects, the irreversible open-system models have been applied to investigations of the effects of phonon scattering, as described in Appendix F, and the self-consistent potential in the RTD, as described in Appendix A. The various applications of open-system kinetic theory to RTD's clearly demonstrate the value of this approach, in spite of the existence of several unresolved mathematical issues which will be explored in the next section.

VI. PROPERTIES OF THE IRREVERSIBLE MODEL

A. Mathematical properties

Having demonstrated the computational utility of the time-irreversible open-system model defined by Eqs. (4.4) and (4.7), let us examine its properties in more detail. First, note that the Wigner function derived from a steady-state (4.27) or transient solution of Eq. (4.4) is purely real valued, because both the Liouville equation (4.4) and the boundary conditions (4.7) are purely real. This implies that the corresponding density matrix is Hermitian, as required.

Now consider the domain upon which the model is defined, as contrasted to the domain of a spatially closed system. This is illustrated in Fig. 20. For a closed system of length l (bounded by an infinite potential well), the state of the system would be described by a density matrix defined within the square formed by the long-dashed lines. The coordinate rotation from the Wigner-Weyl transformation (4.1) implies that the domain of the Wigner function maps onto the rotated square ("diamond-shaped domain") shown by the short-dashed lines in the x,x' plane. The density operator is, in effect, a spatial correlation function. The partitioning of a one-dimensional "universe" into a finite system bounded by two semi-infinite reservoirs partitions the domain of the density operator into regions corresponding to various system-system, system-reservoir, and reservoir-reservoir correlations. The domain of the Wigner function does not coincide with that of the system-system density operator, and the Wigner function domain extends into regions that describe system-reservoir correlations. This may well be a necessary characteristic of any useful open-system model.

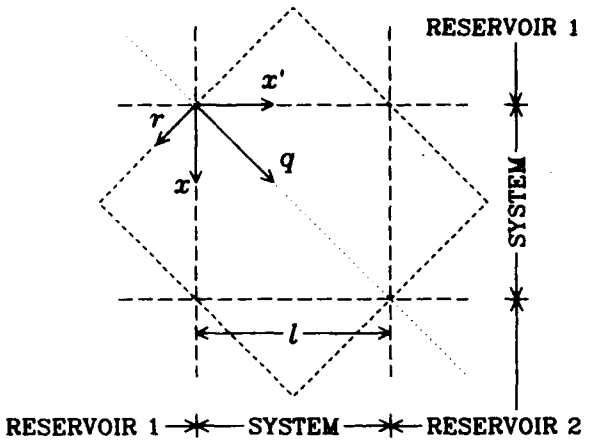


FIG. 20. Domain of the density matrix and the Wigner distribution function. The arguments of the density matrix are x and x'. The Wigner function is obtained by transforming to the coordinates q and r, followed by a Fourier transform with respect to r. The long-dashed lines indicate the system-reservoir boundaries, and they partition the domain into regions corresponding to the various system-system, system-reservoir, and reservoir-reservoir correlations. The short-dashed lines represent the boundaries of the domain of the Wigner-distribution-function model. Note that the Wigner function includes contributions from regions that represent correlations with the reservoirs.

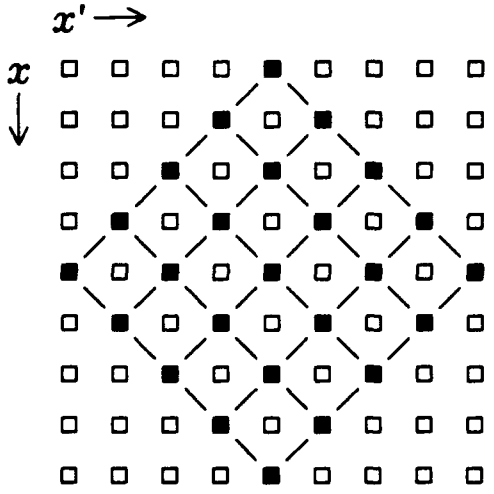


FIG. 21. Illustration of the inconsistency between discretizations for the density operator and the Wigner function. The squares represent the elements of a bounded, discrete density operator. To transform this into a Wigner function only the filled squares may be employed because they form a discrete, rectangular mesh in the (q,r) space. This not only leaves the elements in the corner triangles of the density operator unused, but employs only one-half of the remaining elements. As a result, the transformation from discrete density operator to discrete Wigner function is not unitary.

It must be admitted that the shape of the Wigner-function domain as shown in Fig. 20 introduces certain mathematical difficulties. These arise when one requires the density operator given the Wigner function and vice versa. First let us note that the Wigner-Weyl transformation of the density operator into the Wigner function is a unitary superoperator in the sense of Eq. (2.6) if the domain [in (x,x') and (q,p)] is unbounded. This follows from the equivalence of the inner products (2.4) and (4.23). If the domains in (x,x') and (q,r) are bounded and do not coincide, the Wigner-Weyl transformation cannot be unitary (and is in fact noninvertible), because some of the information contained in either the Wigner function or the density operator will be lost. This is precisely the situation illustrated in Fig. 20. An additional problem arises in the discrete model which involves the form of the discrete mesh in the two coordinate systems. This is illustrated in Fig. 21, which shows a discrete mesh in (x,x') and superimposed upon it the rectangular mesh in (q,r) employed in Eq. (4.13). In addition to the loss of information from the corner triangles described above, there is also a loss of information because the (q,r) mesh points are only half as dense as the (x,x') mesh points. The relation between these two meshes can be summarized as $\Delta_q=\Delta_x$ and $\Delta_r=2\Delta_x$. [This mesh is implicitly used in Eq. (4.14).] If the (q,r) mesh were set up with $\Delta_q=\Delta_x$ and $\Delta_r=\Delta_x$, half of the (q,r) mesh points would not coincide with the (x,x') points. A way to incorporate all the (x,x') points might be to use a staggered mesh in (q,r) with $\Delta_q=\frac{1}{2}\Delta_x$ and $\Delta_r=2\Delta_x$. Mains and Haddad (1989) have investigated such a scheme.

In summary, one cannot rigorously derive a Wigner function from a density operator and vice versa on a finite, and particularly on a discrete, domain. As a result, any discussions that rely upon the equivalence between the Wigner function and the density operator in such a case must be regarded as plausibility arguments rather than derivations. A more practical consequence is that we have no adequate way to evaluate the operator properties, such as the eigenvalue spectrum or the inverse, of a Wigner function defined upon a bounded domain.

The shape of the natural domain for the Wigner function is a consequence of its relationship with the superoperators generated by x and $p_x=(\hbar/i)\partial/\partial x$. In terms of the variables q, p, and r, these superoperators have particularly simple forms:

$$\mathcal{X}_{(+)}=q\ , \tag{6.1}$$

$$\mathcal{X}_{(-)}=r=i\hbar\frac{\partial}{\partial p}\ , \tag{6.2}$$

$$\mathcal{P}_{(+)}=p=\frac{\hbar}{i}\frac{\partial}{\partial r}\ , \tag{6.3}$$

$$\mathcal{P}_{(-)}=\frac{\hbar}{i}\frac{\partial}{\partial q}\ . \tag{6.4}$$

The Wigner function is thus expressed in terms of the eigenvalues of $\mathcal{X}_{(+)}$ and $\mathcal{P}_{(+)}$, and the fact that these superoperators commute [Eq. (2.13)] is what allows us to define the Wigner function in the first place (because its

arguments are the eigenvalues of these superoperators). This observation is the point from which to begin to address one of the obvious concerns connected with any phase-space formulation of a quantum problem: the possibility of a violation of the uncertainty principle. Because q and p are eigenvalues of commuting superoperators, specifying boundary values localized in the (q,p) plane does not necessarily lead to a violation of the uncertainty principle.

How, then, does the uncertainty principle affect the Wigner function? The usual characteristic of a distribution functon that violates the uncertainty principle is that it contains some states which have negative occupation probabilities. That is, the corresponding density matrix will have at least some negative eigenvalues. Consider, for example, a distribution function $f(q,p)=\pi\delta(q)\delta(p)$, which clearly violates the uncertainty principle. The corresponding density matrix is $\rho(x,x')=\delta(x+x')$. If we operate on any antisymmetric state $\psi_a(x)=-\psi_a(-x)$ with this density matrix, we get $-\psi_a(x)$, so -1 is certainly an eigenvalue of ρ, which is thus not a valid density matrix. [Note, however, that examples of distribution functions that satisfy the uncertainty principle and are still not valid Wigner functions have been found (Narcowich and O'Connell, 1986)].

Therefore, to represent an acceptable mixed state, the density operator ρ must be a positive operator. (Recall that we have modified the normalization condition so that $\mathrm{Tr}\rho=1$ is no longer a requirement.) The positivity of ρ and thus of f as an operator does *not* imply that $f(q,p)\geq 0$. It is well known that the Wigner function can take negative values (Wigner, 1971), and that such negative values are related to quantum interference, as we have seen. One can test the positivity of ρ using two different conditions (Narcowich and O'Connell, 1986). The most commonly invoked approach is to demand that

$$\langle\psi|\rho|\psi\rangle\geq 0\ , \tag{6.5}$$

for all states ψ. The expectation value can be rewritten as an operator inner product [Eq. (2.4)] by defining the projection operator $P_\psi=|\psi\rangle\langle\psi|$:

$$\langle\psi|\rho|\psi\rangle=\mathrm{Tr}(P_\psi\rho)=\langle P_\psi\|\rho\rangle\ . \tag{6.6}$$

Then the condition (6.5) can be transformed into the Wigner-Weyl representation using Eq. (4.23) to obtain the condition

$$\int dq\int dp\, f(q,p)f_\psi(q,p)\geq 0\ , \tag{6.7}$$

(where f_ψ is the Wigner function for the pure state ψ) for all ψ. The application of this condition to the distribution functions obtained from the open-system model is hindered by the problems of incompatibility of the finite domains discussed above. In the second test for positivity of the density operator one demands that it be possible to factor ρ into

$$\rho=A^\dagger A\ , \tag{6.8}$$

where A is some operator (Narcowich and O'Connell, 1986). Applying this condition to the corresponding distribution function requires the expression for the operator product in terms of Wigner functions (Hillery, O'Connell, Scully, and Wigner, 1984). Condition (6.8) then becomes (Narcowich and O'Connell, 1986)

$$f(q,p)=\frac{1}{(\pi\hbar)^2}\int dq'\int dp'\int dq''\int dp''a^*(q+q',p+p')a(q+q'',p+q'')e^{2i(q'p''-q''p')/\hbar}\ , \tag{6.9}$$

where $a(q,p)$ is the Wigner-Weyl transform of A. It appears that the obvious ways to restrict the limits of integration in Eq. (6.9) to a finite domain lead to expressions that violate at least one of the semi-group axioms which define operator multiplication. If an expression that did satisfy those axioms could be derived from Eq. (6.9), we would obtain a useful definition of positivity in the open-system case.

Now, does the procedure of directly solving for the Wigner function under inhomogeneous boundary conditions lead to a positive $f^{(\mathrm{dc})}$ operator? In the absence of a rigorous definition of positivity for a Wigner function on a finite domain, there is, of course, no mathematical demonstration that guarantees such positivity. It may well be possible to define a case of the present open-system model which does violate the uncertainty principle. However, let us qualitatively explore some of the considerations that bear upon this question. First, note that the positivity of $f^{(\mathrm{dc})}$ necessarily involves the positivity of the boundary values, because $f^{(\mathrm{dc})}$ is a linear function of the boundary values as shown by Eqs. (4.27). We can speculate that at least in a semiclassical situation $f^{(\mathrm{dc})}$ should be a positive operator if $f^{(\mathrm{left})}_{\mathrm{boundary}}$ and $f^{(\mathrm{right})}_{\mathrm{boundary}}$ are positive. To establish the plausibility of the idea, let us consider the classical case. The properties of the classical Liouville equation (4.6) employing the open-system boundary conditions (4.7) are essentially the same as those of the quantum case with respect to the eigenvalue spectrum of the Liouville operator and the stability of the resulting solutions. If we assume that there is no damping within the system, then the classical Liouville theorem holds within the system, and the distribution function f_{cl} is constant along the classical trajectories (which are the characteristic curves of the Liouville equation). Any trajectory passing through a boundary must in fact pass through a boundary twice, once as an incoming particle and once as an outgoing particle (otherwise a density would have to build up in violation of the Liouville theorem). Such trajectories cover the phase space, except for those regions which correspond to any bound orbits. Because f_{cl} is constant along a trajectory and its value is fixed by the boundary condition, f_{cl} must be

non-negative if, and only if, the boundary values are non-negative. The values of f_{cl} in regions corresponding to bound states will be non-negative if and only if the initial values of f_{cl} (with respect to time) are non-negative.

How might these considerations be modified in a quantum-mechanical system? Or, in other words, how can one get into trouble applying the open-system boundary conditions to a quantum system? The only obvious case would be an attempt to apply the boundary conditions (4.7) in a region where there were strong interference effects, such as standing waves. We can easily imagine that, for example, forcing f to have a large density at a boundary point where a node in the density should occur would introduce spurious states with negative occupation. To avoid such situations, one should apply Eq. (4.7) only in reasonably classical regions of a system. In practice, this means at a distance of at least a few times the thermal coherence length λ_T [Eq. (3.3)] away from any abrupt feature of the potential (where the standing waves are smeared out by thermal incoherence). At lower temperatures, one would use the reciprocal of the Fermi wave vector, rather than λ_T.

Now let us examine in more detail the mathematical structure of the model that results from the time-irreversible boundary conditions. The discrete expression for the drift term $\mathcal{T}$ of the Liouville equation (4.19) has the form of a master operator (Bedeaux, Lakatos-Lindenberg, and Shuler, 1971). Such an operator, when applied to a distribution function, has the effect of removing some fraction of the density in each possible state and redistributing that fraction among the other possible states. For a finite, discrete model the properties of the matrix M representing a master operator are

$$\begin{aligned} &m_{ii} \le 0 , \\ &m_{ij} \ge 0 \quad \text{for } i \ne j , \\ &\sum_i m_{ij} \le 0 . \end{aligned} \tag{6.10}$$

In the last condition the column sum is actually equal to zero except for those states j which can lose density to an external reservoir, as is the case for the open-system model on the outflowing boundaries. All the eigenvalues of a matrix satisfying the conditions (6.10) will have non-positive real parts (Oppenheim, Shuler, and Weiss, 1977, Chap. 3). This may be readily demonstrated by appealing to Gerschgorin's theorem (Wilkinson, 1965), which states that every eigenvalue of a matrix A lies in at least one of the circular discs (in the complex plane) with centers at a_{ii} and radii $\sum_{i \ne j} |a_{ij}|$. To apply this theorem to the master operator M, let us take the matrix A to be the transpose of M, $A = M^T$, to change the column sum condition into a row sum. The eigenvalues of M and A are identical. Then because a_{ij} is negative for $i = j$ and positive for $i \ne j$ and is real for all i and j, we find that the real part of each eigenvalue λ_k must satisfy

$$a_{ii} - \sum_{j \ne i} a_{ij} \le \mathrm{Re}\lambda_k \le a_{ii} + \sum_{j \ne i} a_{ij} = \sum_j a_{ij} \le 0 , \tag{6.11}$$

for some i. Thus $\mathrm{Re}\lambda_k \le 0$ for all k. The fact that the column sums in $\mathcal{T}$ for the outflow boundaries are less than zero makes $\mathcal{T}$ nonsingular. (In a master operator describing a closed system, all the column sums would be zero, which implies that the determinant would be zero, so there must be an eigenvalue equal to zero.)

The fact that the upwind discretization generates a master operator is the fundamental reason for its success, both in the present context and in the more traditional applications of transport theory (Roache, 1976, pp. 4–5; Duderstadt and Martin, 1979). Now, in the quantum case, the complete Liouville operator $\mathcal{L}$ (in the Wigner-Weyl representation) cannot be a master operator, because we know that the Wigner distribution can have negative values, which a master operator would not permit. As we have noted, the quantum-interference phenomena enter the Wigner distribution via the potential superoperator $\mathcal{V}$. The fundamental result of the present work is the demonstration in Fig. 9 and Eqs. (4.24) and (4.25) that the Markovian model which follows from the irreversible boundary conditions (4.7) introduces the necessary stability properties in the quantum case as well as in the much more obvious classical case.

It is interesting to consider the form that $\mathcal{T}$ assumes upon transformation back to a real-space density-matrix representation. For this purpose let us assume that we have defined the Wigner function on a discrete basis with respect to q and on a continuum basis with respect to p. Then $\mathcal{T}$ is given by

$$(\mathcal{T}f)(q,p) = -\frac{p}{m\Delta_q} \times \begin{cases} f(q+\Delta_q,p) - f(q,p) & \text{for } p < 0 \\ f(q,p) - f(q-\Delta_q,p) & \text{for } p > 0 \end{cases} . \tag{6.12}$$

To transform this back to the density-matrix representation, we must evaluate

$$(\mathcal{T}\rho)(q,r) = \int_{-\infty}^{\infty} \frac{dp}{2\pi\hbar} e^{ipr/\hbar} (\mathcal{T}f)(q,p) , \tag{6.13}$$

with Eq. (4.3) substituted for f. [To simplify the resulting expressions, we shall express the arguments of ρ in terms of q and r of Eq. (4.1) and Fig. 20.] Evaluation of Eq. (6.13) requires the formula

$$\int_0^{\infty} \frac{dp}{2\pi\hbar} p e^{ipr/\hbar} = \hbar \frac{\partial}{\partial r} \left[\frac{1}{2\pi} \wp \frac{1}{r} - \frac{i}{2} \delta(r) \right] \tag{6.14}$$

and its complex conjugate. Letting Δ_q approach zero we find

$$(\mathcal{T}\rho)(q,r) = \frac{\hbar}{m} \frac{\partial}{\partial r} \left[i \frac{\partial \rho(q,r)}{\partial q} + \frac{\Delta_q}{2\pi} \wp \int_{-\infty}^{\infty} \frac{dr'}{r-r'} \frac{\partial^2 \rho(q,r')}{\partial q^2} \right] . \tag{6.15}$$

The second term in Eq. (6.15) contributes an anti-Hermitian component to $\mathcal{L}$. The appearance of $\partial^2/\partial q^2$ in this term is reminiscent of the "numerical viscosity"

that is a property of some finite-difference formulations of transport equations (Press, Flannery, Teukolsky, and Vetterling, 1986). The principal-value integral in Eq. (6.15) has the desired effect of distinguishing the sign of the momentum of the states present in ρ. To see this, suppose that there is a term $|k\rangle\langle k| = e^{ikr}$ contained in ρ. One could evaluate its contribution to the integral in (6.15) by contour integration, closing the contour in the upper or lower half-plane if k were positive or negative, respectively. But then the sign of the contribution of the pole on the real axis would change as the sign of k changes. The anti-Hermitian term would vanish, except possibly for a surface contribution, in the limit $\Delta_q \to 0$.

This description of open systems in terms of $\rho(x,x')$ has not yet been developed into a workable model. However, there is a strong motivation for doing so in the context of semiconductor heterostructures. In such a structure the electron energy-momentum relation can be considerably more complex than a simple parabola, and it changes from one material to another in ways that cannot be represented by a shift in the local potential. The simplest example of such an effect is the change in effective mass as an electron crosses a heterojunction. As described in Appendix E, this leads to a highly nonlocal form for the kinetic-energy superoperator in the Wigner-Weyl representation. More complex features of the energy-band structure can be modeled by any of a number of localized-basis-function schemes which may requires more than one basis function per lattice site. Such schemes could easily fit into an approach expressed in terms of $\rho(x,x')$, but it is not at all obvious how to incorporate such effects into the Wigner function in view of the incompatible discretization requirements illustrated in Fig. 21.

Of more general interest is the appearance of Eq. (6.14) in the deductive chain leading to (6.15). Such a relation, more often expressed in the form

$$\frac{1}{\omega \mp i\epsilon} = \mathcal{P}\frac{1}{\omega} \pm i\pi\delta(\omega) , \tag{6.16}$$

is usually encountered in the analysis of irreversible quantum phenomena. It is the mathematical expression of the fact that a continuum of states (and therefore of frequencies) provides enough degrees of freedom that a Poincaré recurrence can be postponed indefinitely. It appears in the analysis of behavior in the time and frequency domains, and is used to express the initial conditions that lead to irreversible behavior: no advanced waves in electrodynamics (Bjorken and Drell, 1964), or adiabatic switching-on in many-body theory (Kohn and Luttinger, 1957; Fetter and Walecka, 1971). In the present model such a relation appears in the position and momentum domains and expresses the effects of the spatial boundary conditions.

B. Superoperator symmetry and physical observables

One of the benefits of the time-irreversible open-system boundary conditions is that they provide an alternative to the use of periodic boundary conditions in the analysis of quantum-transport phenomena. The great disadvantage of periodic boundary conditions is that they do not address the case in which the potential varies significantly across a system. That is, their use restricts one to the study of low-field phenomena. It has been pointed out (Yennie, 1987, footnote 11 acknowledging private discussion with M. Weinstein) that quasiperiodic boundary conditions (i.e., periodic within a phase factor which can be removed by a gauge transformation) are necessary if the momentum operator is to be Hermitian on a finite domain. The present work demonstrates that far-from-equilibrium phenomena can be modeled by employing a non-Hermitian momentum superoperator.

The connection between symmetries and conservation laws is undoubtedly one of the most fundamental results of the quantum theory. However, if one is faced with the task of describing the behavior of a nonconservative system, the inability to modify or violate the conservation laws becomes an obstacle to defining a realistic model, rather than a benefit. The problem is that one wants a model whose solutions stably approach a steady state, which requires complex-valued eigenvalues, but the expectation values of physical observables should be real. The present analysis of open-system models demonstrates that these conflicting requirements can be accommodated at the kinetic level, because the roles of generating the dynamical evolution and evaluating observables are filled by different *superoperators*. If we reexamine the models described above, we find that the dynamic effects such as generating time evolution or moving density by current flow are described by commutator superoperators, and these are the superoperators that become non-Hermitian when one incorporates interactions with the outside world. The measurement of the expectation values of observables is done by anticommutator superoperators, and these, with proper attention to the definition of the domain and boundary conditions, remain Hermitian. This separation of function has been noted by Prigogine (1980) in the superoperators generated by the Hamiltonian. In the open-system model the momentum superoperators appear in similar roles, and this demonstrates the existence of a more general underlying structure in the kinetic theory.

Let us consider the superoperators $\mathcal{P}_{(+)}$ and $\mathcal{P}_{(-)}$ derived from the momentum operator. We have already observed that the kinetic-energy term of the Liouville equation (2.3) can be written as $\mathcal{P}_{(+)}\mathcal{P}_{(-)}/m$ (3.6). $\mathcal{P}_{(+)}$ will be Hermitian if we restrict our attention to density matrices whose off-diagonal elements approach zero for large $x - x'$ (so that integration by parts may be performed without a surface contribution in an integral over $r = x - x'$). Such density matrices describe normal systems (as opposed to superconducting ones, or systems with some other long-range coherent effect) at nonzero temperature. In such normal cases $\mathcal{P}_{(+)}$ produces the real-valued factor p in the drift term (4.9). $\mathcal{P}_{(-)}$ generates the gradient in $\mathcal{T}$ and is thus the superoperator that is

rendered non-Hermitian by the boundary conditions (4.7).

We can also see the dichotomy of function between $\mathcal{P}_{(+)}$ and $\mathcal{P}_{(-)}$ by examining the elementary quantum continuity equation, which is conventionally written

$$\frac{\partial}{\partial t}\psi^*\psi = -\frac{\partial}{\partial x}(\psi^* J \psi) , \tag{6.17}$$

where

$$(\psi^* J \psi) = \frac{\hbar}{2mi}\left[\psi^*\frac{\partial\psi}{\partial x} - \frac{\partial\psi^*}{\partial x}\psi\right] . \tag{6.18}$$

By now we should readily recognize the presence of $\mathcal{P}_{(+)}$ in the current-density operator J. In fact, the current density is much more naturally regarded as a superoperator,

$$\mathcal{J} = \mathcal{P}_{(+)}/m , \tag{6.19}$$

and we see $\mathcal{P}_{(+)}$ in the role of measuring an observable. At the kinetic level the continuity equation is linear in terms of the density matrix ρ and is simply the Liouville equation evaluated along the diagonal $x = x'$.

The continuity equation is of course just the zeroth-order moment equation of the Liouville equation. The higher-order moments of the Liouville equation may be obtained by operating on the equation with $\mathcal{P}_{(+)}$ (or $\mathcal{J}$) and evaluating the resulting expression along the diagonal. Let us denote the evaluation of an operator kernel for $x = x' = q$ by angular brackets, $\langle\rho\rangle_q = \rho(q,q)$. This is equivalent to the phase-space procedure of multiplying by some power of p and then integrating over all p, so that the corresponding expression for the Wigner function is

$$\langle \mathcal{P}_{(+)}^n \rho \rangle_q = \int_{-\infty}^{\infty} \frac{dp}{2\pi\hbar} p^n f(q,p) . \tag{6.20}$$

The moment equations we shall derive are a special case of those that have been discussed by a number of authors (Frölich, 1967; Putterman, 1974; Iafrate, Grubin, and Ferry, 1981; Kreuzer, 1981), because we shall not consider two-body or dissipative interactions. The objective is to demonstrate the role of the anticommutator superoperators in this procedure, a point that has not been previously articulated.

As a starting point from which to derive the moment equations, let us rewrite the Liouville equation in superoperator notation, making use of the factorization (3.6):

$$\frac{\partial\rho}{\partial t} = \frac{1}{i\hbar}\left[\frac{1}{m}\mathcal{P}_{(-)}\mathcal{P}_{(+)} + \mathcal{V}_{(-)}\right]\rho = -\frac{1}{m}\frac{\partial}{\partial q}\mathcal{P}_{(+)}\rho + \frac{1}{i\hbar}\mathcal{V}_{(-)}\rho . \tag{6.21}$$

The manipulations required to generate the moment equations may be considerably simplified by using some superoperator relations to evaluate the effect of $\mathcal{P}_{(+)}$ on the potential and its derivatives. To derive the necessary expressions, let us consider an operator $G = g(x)\delta(x - x')$, which is diagonal in position space. The commutators of the derived superoperators $\mathcal{G}_{(-)}$ and $\mathcal{G}_{(+)}$ with $\mathcal{P}_{(+)}$ are then

$$\begin{aligned}[\mathcal{P}_{(+)}, \mathcal{G}_{(-)}] &= -i\hbar\mathcal{G}'_{(+)} , \\ [\mathcal{P}_{(+)}, \mathcal{G}_{(-)}] &= -\tfrac{1}{4}i\hbar\mathcal{G}'_{(-)} ,\end{aligned} \tag{6.22}$$

where $\mathcal{G}'$ indicates the superoperator derived from the spatial derivative $\partial g/\partial x$. Note that $\langle \mathcal{G}_{(-)}\rho\rangle_q = 0$ for any such operator, and $\langle \mathcal{G}_{(+)}\rho\rangle_q = g(q)\langle\rho\rangle_q$ for any operator ρ. Now we may readily derive the moment equations. The zeroth moment is thus

$$\frac{\partial\langle\rho\rangle_q}{\partial t} = -\frac{\partial\langle\mathcal{J}\rho\rangle_q}{\partial q} , \tag{6.23}$$

which is a familiar form of the continuity equation. If a collision term is included in the kinetic equation, it must have a form such that $\langle \mathcal{C}\rho\rangle_q = 0$ if the theory is to satisfy the continuity equation. This means that $C(x_1,x_1;x_2,x_2') = 0$ in the density-matrix representation [a condition satisfied by the Fokker-Planck operator (3.8)] or

$$C(q_1,p_1;q_2,p_2) = \delta(q_1 - q_2)c(p_1,p_2,q_1)$$

with $\int c(p_1,p_2,q)dp_1 = 0$, for the Wigner function (see Appendix F).

The first moment equation is readily found to be

$$m\frac{\partial\langle\mathcal{J}\rho\rangle_q}{\partial t} = -\frac{\partial}{\partial q}\langle\Pi\rho\rangle_q - \frac{\partial v}{\partial q}\langle\rho\rangle_q , \tag{6.24}$$

where $\Pi = \mathcal{P}_{(+)}^2/m$ is the momentum flux density. [For two- or three-dimensional models, the direct product of the two vectors $\mathcal{P}_{(+)}$ is taken, and Π will be a tensor (Landau and Lifshitz, 1959).] Equation (6.24) is identical to its classical counterpart. If we integrate it with respect to q [assuming that the domain is rectangular in the (q,r) coordinates and extends over $0 < q < l$], we obtain a generalization of Ehrenfest's theorem to the case of an open system:

$$m\frac{\partial}{\partial t}\int_0^l\langle\mathcal{J}\rho\rangle_q dq = -\int_0^l\frac{\partial v}{\partial q}\langle\rho\rangle_q dq + \langle\Pi\rho\rangle_0 - \langle\Pi\rho\rangle_l . \tag{6.25}$$

The last two terms represent the effect of opening the system: A flux of momentum density through the boundaries of the system will affect the current flow within the system. To make contact with hydrodynamics, we would follow the standard kinetic-theory manipulations (Kreuzer, 1981, Chap. 8) and define a kinetic pressure tensor

$$\mathbf{P} = \langle(\mathcal{P}_{(+)} - \langle\mathcal{P}_{(+)}\rangle)^2\rangle/m$$

and separate Π into

$$\langle \Pi\rho \rangle_q = \langle \mathbf{P}\rho \rangle_q + \langle \mathcal{P}_{(+)}\rho \rangle_q^2 / m \ .$$

Continuing with the above procedure, we may derive a second moment equation by operating on Eq. (6.21) with $\mathcal{P}_{(+)}^2/2m$ to obtain

$$\frac{\partial}{\partial t}\left\langle \frac{\mathcal{P}_{(+)}^2}{2m}\rho \right\rangle_q = -\frac{\partial}{\partial q}\left\langle \frac{\mathcal{P}_{(+)}^3}{2m^2}\rho \right\rangle_q - \frac{\partial v}{\partial q}\langle \mathcal{J}\rho \rangle_q \ . \tag{6.26}$$

Quantum corrections in the form of terms containing $\hbar^2\partial^3 v/\partial q^3$ will begin to appear in the third and higher moment equations, as one would expect from the Wigner-Moyal expansion (4.12). However, the second moment equation presents something of an ambiguity. We might also derive it by operating on Eq. (6.21) with the $\mathcal{T}_{(+)}$ derived from the kinetic-energy operator T. These are not at all the same superoperators:

$$\frac{\mathcal{P}_{(+)}^2}{2m} = -\frac{\hbar^2}{8m}\left[\frac{\partial^2}{\partial x^2} - 2\frac{\partial^2}{\partial x \partial x'} + \frac{\partial^2}{\partial x'^2}\right] , \tag{6.27}$$

$$\mathcal{T}_{(+)} = -\frac{\hbar^2}{4m}\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial x'^2}\right] . \tag{6.28}$$

Putterman (1974) displays both of these forms and notes that both lead to the same bulk properties, thus any physical difference must appear in a surface contribution. It is not the purpose of the present discussion to investigate these issues in detail, but only to demonstrate that anticommutator superoperators appear naturally in any attempt to evaluate expectation values in kinetic theory.

The same dichotomy between commutator and anticommutator superoperators can be seen in the case of the superoperators generated by the Hamiltonian H. Of course $\mathcal{H}_{(-)}$ is just the Liouville superoperator $\mathcal{L}$, and we have examined at length the need for a departure from Hermiticity in the case of $\mathcal{L}$. We have not yet encountered a need for the anticommutator $\mathcal{H}_{(+)}$. One place it does occur is in a generalization of the Bloch equation (3.1) to the case of an open system. If one attempts to compute an equilibrium density matrix as a finite segment of a much larger system by modifying the boundary conditions on ρ in the Bloch equation, one quickly discovers that product $H\rho$ must be symmetrized to obtain sensible answers. Thus the Bloch equation becomes

$$\partial\rho_{eq}/\partial\beta = -\tfrac{1}{2}(H\rho_{eq} + \rho_{eq}H) = -\mathcal{H}_{(+)}\rho_{eq} \ . \tag{6.29}$$

If the time-reversible open-system boundary conditions (3.4) are applied to the Bloch equation, one obtains a quite useful method for evaluating the equilibrium density matrix (in contrast to the disastrous effect these boundary conditions have upon the time evolution). Taking into account our particle-density normalization of ρ, we find that the correct Bloch equation is

$$\partial\rho_{eq}/\partial\beta = -(\mathcal{H}_{(+)} - \mu)\rho_{eq} \ , \tag{6.30}$$

with the initial condition

$$\rho_{eq}|_{\beta=0} = \delta(x - x') \ . \tag{6.31}$$

When this equation is integrated, the resulting densities in regions of constant potential are found to be equal to the semiclassically expected value

$$(\sqrt{2\pi}\lambda_T)^{-1}\exp[\beta(\mu - v)] \ .$$

An example of an equilibrium density matrix obtained from such a calculation is shown in Fig. 22.

Here we see that again the anticommutator superoperator appears in the process of evaluating an observable, in this case for the purpose of evaluating the energy and thus the occupation probability of the possible states. We would expect that, for this purpose, $\mathcal{H}_{(+)}$ ought to be Hermitian. Its Hermiticity in fact depends upon the shape of the domain when the boundary conditions (3.4) are applied. Because $\mathcal{H}_{(+)}$ is an elliptic operator, it is easy to show that it will be Hermitian when the domain is rectangular in the (q,r) coordinates, so that the gradient in (3.4) is normal to the system boundary. It is not Hermitian when applied to a domain that is square in the (x,x') coordinates, as in the calculation illustrated in Fig. 22. However, the departure from Hermiticity is small,

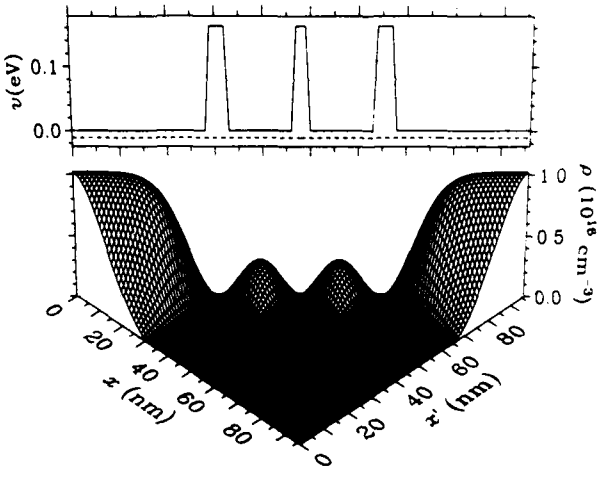


FIG. 22. Equilibrium density matrix obtained by numerically integrating the generalized Bloch equation (6.30) subject to the reversible open-system boundary conditions (3.4). The potential, displayed above, represents the sort of features that are now realizable using semiconductor heterostructure technology. The chemical potential μ is indicated by the dashed line. The calculation employed parameters appropriate for the $Al_xGa_{1-x}As$ system at 77 K. The three energy barriers create two identical "quantum wells," bounded by contacting layers. The lowest energy states in these wells are pushed toward higher energy by size quantization, which reduces the electron density in the wells via the Boltzmann factor. The shallow peaks off the diagonal measure the correlation between the phase of the electron at different positions, and indicate in the present case that the symmetric combination of the well states has a greater occupation factor than the antisymmetric combination.

and the results are physically quite reasonable. I have not yet implemented a program to perform such a calculation on a rectangular domain in (q,r), but this would be the proper way to proceed to evaluate the equilibrium density matrix using open-system boundary conditions.

Having noted that $\mathcal{H}_{(+)}$ appears in the evaluation of the equilibrium density matrix, we can address a point raised by Dahl (1981). It is that $\mathcal{L}$, by itself, does not define a unique eigenvalue problem *in the wave-function space* of a quantum system; but together with $\mathcal{H}_{(+)}$, it does define such a problem. This consideration enters the present problem only for bound states localized within the open system (Carruthers and Zachariasen, 1983). As noted earlier, such states would lead to a nontrivial null space of $\mathcal{L}$. The occupation of such states would have to be determined as an initial condition, such as an equilibrium distribution evaluated using $\mathcal{H}_{(+)}$.

C. Relation to many-body theory

I have remarked that the Markovian kinetic models considered here are not equivalent to the usual elementary quantum-mechanical models of systems such as tunneling diodes. Let us now explore the differences between these two types of models by examining how they may be viewed as different approximations to a single many-body theory. In the approach to many-body transport theory developed by Kadanoff and Baym (1962) and by Keldysh (1964) and elaborated by Langreth (1976) and by Mahan (1987), the description of a quantum system is contained in a Green's function,

$$G^{<}(x,t;x',t')=i\langle\Psi^{\dagger}(x',t')\Psi(x,t)\rangle\ , \tag{6.32}$$

where Ψ is the field operator. The density operator ρ can be obtained from

$$\rho(x,x';t)=-iG^{<}(x,t;x',t)\ . \tag{6.33}$$

Note, however, that the Green's function has, in general, a second time argument t', and this supplies the additional degree of freedom required to describe non-Markovian behavior. The demonstration of the correspondence between the Green's-function formalism and more classical transport equations proceeds applying a Wigner-Weyl-like transformation to the time variables: Define new variables $T=\frac{1}{2}(t+t')$ and $\tau=t-t'$, and then Fourier transform $G^{<}$ with respect to τ:

$$G^{<}(x,x',t,\omega)=\int dt\, e^{i\omega\tau}G^{<}(x,x',T,\tau)\ . \tag{6.34}$$

In the absence of interactions, the equations of motion for $G^{<}$ then become (Mahan, 1987), in the present notation,

$$\left[i\hbar\frac{\partial}{\partial T}-\mathcal{L}\right]G^{<}=0\ , \tag{6.35}$$

$$(\hbar\omega-\mathcal{H}_{(+)})G^{<}=0\ . \tag{6.36}$$

[If interactions are present, collision terms involving the self-energy appear on the right-hand side of Eqs. (6.35) and (6.36).] Without interactions, Eq. (6.35) is just the Liouville equation and (6.36) is a symmetrized Schrödinger equation. On an unbounded domain, these equations simply reproduce pure-state quantum mechanics, as noted above, and the usual tunneling theory follows. However, if we restrict the domain so as to obtain the open-system case, and we wish to reproduce the tunneling theory, we have to apply traveling-wave boundary conditions such as those discussed in Appendix D. Such boundary conditions necessarily introduce a dependence upon ω into Eq. (6.35). Even though we are still considering a "noninteracting" system (in the usual sense of no dissipation), we see that additional ω-dependent boundary terms must appear in Eqs. (6.35) and (6.36).

The Markovian models neglect this ω dependence. They are thus not equivalent to the tunneling or scattering theory. One can view such models either as an approximation to the tunneling theory, or alternatively, as simply a different approximation to the underlying many-body theory. In the latter view, the steady-state tunneling theory is obtained by neglecting the T dependence of $G^{<}$, whereas the Markovian model is obtained by neglecting the ω dependence of $G^{<}$. Thus we may regard the Markov approximation as an *a priori* assumption that $G^{<}$ is independent of ω. Inverting the Fourier transform (6.34) shows that this is equivalent to assuming

$$G^{<}(x,x',T,\tau)\stackrel{?}{=}i\rho(x,x';T)\delta(\tau)\ . \tag{6.37}$$

This makes explicit the Markov assumption that the evolution of the system does not depend upon its past history.

To establish the plausibility of the Markov assumption [Eq. (6.37)], let us again consider the picture of an open system as a finite segment of length l of a much larger "universe" of length L which is occupied by a free-electron gas. The Green's function for this noninteracting system is

$$G^{<}(k,\tau)=w_k e^{-iE_k\tau/\hbar}\ , \tag{6.38}$$

where w_k is the probability that state k is occupied and E_k is the energy of that state. Now, by examining $G^{<}$ within the system itself (that is, over $0\le x\le l$ and $0\le x'\le l$) we cannot resolve the wave vectors of any excitations to an accuracy better than $\pm\pi/l$. On the other hand, because the "universe" is of a much larger length L, there will actually be many wave-vector states within any such interval. Thus the $G^{<}$ that one would observe within the system would be an average over these states of the form

$$\bar{G}^{<}(k,\tau)=\frac{L}{2\pi}\int_{k-\pi/l}^{k+\pi/l}dk'w_{k'}e^{-iE_{k'}\tau/\hbar}\ . \tag{6.39}$$

Using $dE/dk=\hbar s_k$, where s_k is the velocity of state k, we can change the integration variable to an energy, and

perform the integral to obtain

$$\bar{G}^{<}(k,\tau)=w_k e^{-iE_k\tau/\hbar}\left|\frac{\sin(\pi s_k\tau/l)}{\pi s_k\tau/L}\right| . \quad (6.40)$$

The bracketed factor approaches $\delta(\tau)$ as $l\to 0$. Now, l is fixed, of course, and thus the width of the "δ function" is fixed. Moreover, the width is just the transit time across the system at the given k. This suggests the interpretation of Eq. (6.40): Any excitation within the system will propagate away (out of the system), and thus its temporal correlation function will decay after a time of the order of the transit time across the system. This demonstrates the motivation for the Markov assumption [Eq. (6.37)] and also its limitation. The generalization of the present open-system model beyond the Markov approximation has not yet been attempted and would be an obvious task for the further development of this approach. [The initial steps in this direction might be found in the work of Ringhofer, Ferry, and Kluksdahl (1989), who study the formulation of nonreflecting boundary conditions for the Wigner function. This work, however, is concerned primarily with obtaining local (in space and time) approximations to the rigorously nonlocal problem.]

VII. DESIGN AND ANALYSIS OF DISCRETE NUMERICAL MODELS

The present work employs numerical computation and modeling for a purpose for which it is not often employed: as the primary mode of investigating the structure and consequences of a physical theory. The more traditional mode of investigation is, of course, to maximize the use of analytical mathematics and resort to numerical techniques only when the opportunities for analysis are exhausted, or when it is necessary to evaluate those complicated expressions which express an analytical solution. Any particular approach to describing physical phenomena will be successful only for some subset of these phenomena and will be otherwise ineffective. Because analytical mathematics is such a widely used tool, its domain of success has been extensively explored; this domain consists of those problems with sufficient symmetry to admit analytic solutions and those problems which can be regarded as small perturbations on analytically soluble problems. For statistical phenomena this generally means thermal equilibrium of analytically tractable systems and very small departures from equilibrium. Numerical simulation techniques that are inherently nonperturbative are better able to address more complex structures and/or far-from-equilibrium states. Because the study of discrete numerical models is not widely practiced, it is worth examining the principles by which such models may be constructed, using the present open-system model as an example.

A common point of view is to regard discrete numerical models, such as finite-difference models for partial differential equations, as approximations to the "truth" embodied in the continuum formulation of the problem (for example, Lapidus and Pinder, 1982). Such a discrete model can represent the continuum solution only to within an accuracy proportional to some power of the mesh spacing (or other appropriate measure of the coarseness of the discrete model). This tends to lead one to believe that the physics of the situation can be represented only to a given order of accuracy, so that such expressions as conservation laws (or balance equations) will be satisfied only to that order (see, for example, Aubert, Vaissiere, and Nougier, 1984). A corollary to this view is that higher-order approximations produce better models. Such is often not the case (Press *et al.*, 1986), because higher-order approximations usually admit spurious short-wavelength modes which adversely affect both the stability and accuracy of such models.

In fact, a better guiding principle is to seek discrete models that are constructed so as to satisfy exactly the physical laws that govern the behavior of the real system. In practice, one often finds that it is possible to satisfy only some, but not all of these laws. *Which* laws are exactly satisfied and the order of the error terms in the remaining laws depend upon the details of the particular discretization scheme. This situation has led to the conventional wisdom that the discretization of partial differential equations is "an art as much as a science" (Press *et al.*, 1986). The science that is often lacking is a consistent analysis of the degree to which all reasonable discretization schemes satisfy the appropriate laws, or preferably the identification of one scheme that exactly satisfies the relevant laws. A particularly attractive example of the latter situation has been given by Visscher (1988, 1989). It is a discretization of Maxwell's equations in three dimensions, which exactly satisfies the integral forms of the equations. This is accomplished by assigning the various field quantities (charge and current density, electric and magnetic field) appropriately to the centers, faces, and edges of cubic finite-difference cells. Unfortunately, we shall see that this ideal situation is not likely to apply to kinetic open-system models, and some trade-offs must be made between the different laws that we wish to satisfy.

A systematic way to determine the advantages and limitations of a discrete model is first to identify the physical laws that the model ought to satisfy and then to evaluate the order of the errors by which the discrete model fails to satisfy those laws. For the present open-system model, I assert that there are four such laws: (i) charge continuity, (ii) momentum balance, (iii) detailed balance of the equilibrium state, and (iv) stability of nonequilibrium states. Energy balance is not included in this list because it adds no physics that is not already described by momentum balance so long as we neglect energy-redistributing processes such as electron-electron or electron-phonon scattering. Condition (iv) is just the criterion that we have examined extensively, that none of

the eigenvalues of the Liouville operator should have a positive imaginary part.

A. Continuity equation

To begin the analysis of the irreversible open-system model defined by Eq. (4.22), let us consider the continuity equation (6.23). First define the discrete approximation to the local particle density $n(q)=\langle\rho\rangle_q$ in the obvious way, converting the integral in (6.20) to a sum:

$$n_j \equiv n(q_j)=\frac{\Delta_p}{2\pi\hbar}\sum_k f_{jk} \ . \tag{7.1}$$

In a discrete model the current density is most naturally regarded as a quantity that is defined on each *interval* between adjacent mesh points, rather than on the mesh points themselves. Thus the divergence of the current density is a difference taken between adjacent intervals and is associated with their common mesh point. Let us denote the current on the interval between q_j and q_{j+1} by $J_{j+1/2}$. Then if J is to satisfy a discrete continuity equation exactly we must define $J_{j+1/2}$ to be

$$J_{j+1/2}=\frac{\Delta_p}{2\pi\hbar}\left[\sum_{k|p_k<0}\frac{p_k}{m}f_{j+1,k}+\sum_{k|p_k>0}\frac{p_k}{m}f_{j,k}\right] . \tag{7.2}$$

The moment of the Liouville equation becomes

$$\frac{\partial n_j}{\partial t}=-\frac{1}{\Delta_q}(J_{j+1/2}-J_{j-1/2}) -\frac{\Delta_p}{2\pi\hbar^2}\sum_{k=1}^{N_p}\sum_{k'=1}^{N_p}V_{j;k,k'}f_{q,k'} \ . \tag{7.3}$$

To show that the contribution from the potential operator V vanishes, let us consider the sum over k first. The sum can be reordered and then $V_{j,k}$ can be expanded using Eq. (4.14):

$$\sum_{k=1}^{N_p}V_{j;k,k'}=\sum_{k=1}^{N_p}V_{j,k}=\frac{2}{N_p}\sum_{j'=1}^{N_q/2}\sum_{k=1}^{N_p}\sin\left[\frac{2k\Delta_p j'\Delta_q}{\hbar}\right](v_{j+j'}-v_{j-j'}) \ . \tag{7.4}$$

Now, this sum will vanish if

$$\sum_{k=1}^{N_p}\sin\left[\frac{2k\Delta_p j'\Delta_q}{\hbar}\right]=0 \ , \tag{7.5}$$

which happens if $(2N_p\Delta_p\Delta_q)/\hbar=2\pi$, and Δ_p was defined so as to satisfy this relation. This is the Fourier completeness relation mentioned earlier. Thus the discrete model exactly satisfies the continuity equation

$$\frac{\partial n_j}{\partial t}=-\frac{1}{\Delta_q}(J_{j+1/2}-J_{j-1/2}) \ . \tag{7.6}$$

The only limit on the precision of this relationship is the arithmetic roundoff error, which is generally several orders of magnitude smaller than typical discretization errors.

Satisfying the continuity equation via the Fourier completeness relation (7.5) relies upon the special properties of the (artificial) Brillouin zone created by the q discretization. To see this, consider k and k' such that $|k-k'|>N_p$. The term $V_{j;k,k'}$ should describe the effect of a short-wavelength component, but because of the ambiguity introduced by the discretization the term is really derived from the much-longer-wavelength component indexed by $(k-k')\mathrm{mod}N_p$. Such an effect is called "aliasing" in the context of signal processing and sampling theory (Oppenheim and Schafer, 1975, Sec. 1.7), where it is generally regarded as undesirable, and it is mathematically the same as an "umklapp process" in the context of solid-state physics. The derivation of the continuity equation in the continuum case relies on no such property; it follows directly from the antisymmetry of the potential kernel V. In a finite model, however, we must cut off the sequence of k's at some value, and this will remove some terms that would need to be present in the summations of the second term of Eq. (7.3) in order to make this term exactly vanish by antisymmetry. Thus, if we do not rely upon the Fourier completeness property, the best we can hope for is to satisfy the continuity equation to $O(\Delta_p)$. The error can be made numerically very small by proper choice of the limiting values of p, but, formally, the continuity equation would not be exactly satisfied.

B. Momentum balance

One begins to encounter the limits of a simple discrete model when the momentum balance (first moment) equation (6.24) is considered. To evaluate the rate of change of current density, insert the discrete Liouville equation (4.22) into the definition of $J_{j+1/2}$ (7.2). One then obtains

$$m\frac{\partial J_{j+1/2}}{\partial t}=-\frac{1}{\Delta_q}(\Pi_{j+1}-\Pi_j)-\frac{\Delta_p}{2\pi\hbar^2}\left[\sum_{k|p_k<0}p_k\sum_{k'}V_{j+1;k,k'}f_{j+1,k'}+\sum_{k|p_k>0}p_k\sum_{k'}V_{j;k,k'}f_{j,k'}\right] , \tag{7.7}$$

where

$$\Pi_j = \frac{\Delta_p}{2\pi\hbar}\left[\sum_{k|p_k<0}\frac{p_k^2}{m}f_{j+1,k}+\sum_{k|p_k>0}\frac{p_k^2}{m}f_{j-1,k}\right]. \tag{7.8}$$

Note that the requirements of consistency in the discretization scheme imply that Π_j, which one might expect to depend only upon the values of f at q_j, actually depends upon the values of f at q_{j-1} and q_{j+1}. This sort of spreading over the domain becomes worse as higher moments are considered. It is probably more correct to regard Π as a local function of q and attribute an error of $O(\Delta_q)$ to Eq. (7.7) (because $f_{j+1,k}\approx f_{j,k}+\Delta_q\partial f/\partial q$). Now consider the potential terms in Eq. (7.7). For simplicity, let us neglect the different j indices required by the form of J and simply evaluate

$$\frac{\Delta_p}{2\pi\hbar^2}\sum_{k=1}^{N_p}p_k\sum_{k'=1}^{N_p}V_{j;k,k'}f_{j,k'}=\left[\sum_{j'=1}^{N_q/2}\frac{\pi}{2N_p\Delta_q}\cot\left[\frac{\pi j'}{N_p}\right](v_{j+j'}-v_{j-j'})\right]\left[\frac{\Delta_p}{2\pi\hbar}\sum_{k'=1}^{N_p}f_{j,k'}\right], \tag{7.9}$$

where Eq. (4.14) is again used and the sums reordered as before. Now, in the continuum case [Eq. (6.24)] this expression reduces to $(\partial v/\partial q)n$. The discrete expression (7.9) shows a functional of v (the first bracketed factor) times n. If we consider only the first term of the sum over j' and take $\cot\alpha\approx 1/\alpha$ for small α, we get $(v_{j+1}-v_{j-1})/2\Delta_q$, which is just the centered-difference approximation to $\partial v/\partial q$. However, the other terms of the sum are not negligible. While $\pi j'/N_p$ is small, the higher terms just add in more remote approximations to $\partial v/\partial q$. Of course, $\cot\alpha$ approaches zero much more rapidly than $1/\alpha$ as α approaches $\pi/2$. Thus there is a natural cutoff of these higher terms so long as $j'\lesssim N_p/2$. This helps to explain the significance of the limit of the j' summation of Eq. (4.14). The value of $N_q/2$ was originally chosen for the upper limit of this sum on the purely empirical basis that the results were most credible with this value, and multiples of N_q were investigated because the summation is carried out in position space. However, most calculations have taken $N_p\approx N_q$, so these conditions are approximately equivalent. The significant result is that the momentum balance equation (6.24) is not satisfied exactly by the discrete model.

The conformance of the discrete model to the momentum balance equation can be significantly improved by modifying the form of the discrete potential operator (4.14). However, this must be approached with some care. One could, for example, simply discretize the classical form $F\partial/\partial p$, and if this is done properly, momentum balance will be exactly satisfied. The problem with this approach, of course, is that it discards any quantum-interference effects. Mains and Haddad (1988a, 1988c) have suggested a better approach. They recommend an alternative expression for $V_{j;k,k'}$ which leads to a model that exactly satisfies a discrete momentum balance expression. The idea is to weight the expression for $V_{j;k,k'}$ as

$$V^{\mathrm{MH}}_{j;k,k'}=\frac{\sin[2\pi(k-k')/N_p]}{2\pi(k-k')/N_p}V_{j;k,k'}. \tag{7.10}$$

If we now evaluate the first moment of the potential term we find

$$\frac{\Delta_p}{2\pi\hbar^2}\sum_{k=1}^{N_p}p_k\sum_{k'=1}^{N_p}V^{\mathrm{MH}}_{j;k,k'}f_{j,k'}=\frac{v_{j+j'}-v_{j-j'}}{2\Delta_q}\left[\frac{\Delta_p}{2\pi\hbar}\sum_{k'=1}^{N_p}f_{j,k'}\right], \tag{7.11}$$

exactly. The use of a weighting function in momentum space corresponds to a convolution in position space. If we reinterpret Eq. (7.10) as a continuum expression, a bit of manipulation will show that (7.10) can be derived from a "smoothed" potential $v^{\mathrm{MH}}(x)=\int dx' w(x-x')v(x')$, where the convolution function w is just a rectangular pulse on the interval $[-\Delta_q,\Delta_q]$. It can be written as $w(x)=\theta(\Delta_q+x)\theta(\Delta_q-x)/2\Delta_q$, where θ denotes the Heaviside step function. Qualitatively, the effect of this scheme is to smooth out any abrupt change in the potential so that any such change is distributed over at least two mesh intervals. However, the convolution theorem does not hold exactly in the finite, discrete domain of the present problem. One consequence of this is that the discretization based upon Eq. (7.10) does not exactly satisfy the continuity equation via the Fourier completeness relation (7.5), but does so only to $O(\Delta_p)$, as discussed above.

A related idea is to use some form of "data windowing" (Oppenheim and Schafer, 1975, Sec. 11.4) in the evaluation of the discrete potential superoperator. This technique is used in the Fourier analysis of finite sets of sampled data, and in the present context would involve multiplying the $(v_{j+j'}-v_{j-j'})$ factor in Eq. (4.14) by some function of j' which decreases to zero for large j' (the window function). That is, the weighting would be done in real space rather than in k space. The objective of data windowing is to maximize the fidelity of the Fourier spectra derived from a finite set of data to those of a hypothetical infinite data set by minimizing the spurious effects associated with cutting off the data at some finite value. Qualitatively, this would seem to suit the requirements of discrete models of quantum systems. Invoking the idea that $V(q,p)$ encodes the quantum-interference effects, we might also interpret a data windowing procedure as an approximate description of the

continuous loss of coherence as one examines points separated farther apart in a dissipative system. This procedure might provide a way to interpolate between the quantum and classical regimes, whereas the obvious schemes for doing so with the Wigner function, expanding in powers of $\hbar$, are known to fail (Heller, 1976). These are intriguing possibilities, but the effects of data windowing on the present sort of open-system models have not been extensively investigated.

C. Detailed balance

The principle of detailed balance is important in describing the properties of the equilibrium state. In the particular case of electron devices it assures us that the current density is zero when the applied voltage (as measured by the difference in chemical potentials) is zero. The reader may have noticed that the concept of equilibrium has played no part in the development of the present open-system model, and indeed the only place where the chemical potential can appear is in the boundary-condition distribution function. In this context it may not be surprising that the discrete model does not exactly satisfy the detailed-balance condition. This was discovered by Jensen and Buot (1989a), who noticed that if the steady-state $J(V)$ curves were computed for a structure lacking inversion symmetry (having unequal barrier widths), a non-negligible current density was obtained at zero bias. Because it is precisely detailed balance which leads us to expect zero current in equilibrium, the spurious equilibrium current is a measure of the violation of this condition.

Given the observation that the discrete model does not exactly satisfy detailed balance, we should determine whether this is a consequence of the discretization or of the open-system boundary conditions themselves. A simple way to do this is to compute the zero-bias current density for an asymmetric RTD structure using varying mesh spacings Δ_q and Δ_p. This was done for a structure identical to that described in Sec. V, except that the widths of the barriers were 3.4 and 2.3 nm. It was found that $J(0)$ was essentially independent of Δ_p and $J(0)=O(\Delta_q)$, as illustrated in Fig. 23. Thus the violation of detailed balance is entirely a result of the discretization, and the continuum formulation will apparently satisfy the detailed-balance principle.

Let us examine this issue in more detail. To begin, let us see what detailed balance implies about the equilibrium density operator or Wigner function. Because the processes occurring in equilibrium must be reversible, the density operator must equal its time-reversed value $\rho_{eq}=\rho_{eq}^{*}$, or $\rho_{eq}(x,x')$ must be purely real. This implies that the equilibrium Wigner distribution must be a symmetric function of p. Thus an alternative measure of the departure from detailed balance is $\langle[f_{eq}(q,p)-f_{eq}(q,-p)]^2\rangle^{1/2}$. Evaluating this measure for computed f_{eq} with various mesh spacings leads to the same conclusion: the irreversible model violates detailed balance to $O(\Delta_q)$, and the error is independent of Δ_p.

The procedure of solving for the steady-state Wigner function and then examining the scaling properties of various features of those solutions is, in the absence of a well-developed and thoroughly checked mathematical analysis, the most reliable way to address such questions as the departure from detailed balance. However, if one is to compare alternative discretization schemes for a particular problem, as is attempted below, it is much more desirable to be able to determine the order of the errors from a knowledge only of the equations (as was done with the moment equations), rather than the solutions. In particular, we want to be able to examine a discretization of the Liouville superoperator and determine the order of error in detailed balance. At present, no simple criterion has been identified that would permit such an analysis, However, we may again examine the factors that bear upon this problem.

Let us again consider the purely classical example of an open system with no internal dissipation. Then the particles will follow their classical trajectories $[q(t),p(t)]$, and along those trajectories the distribution function f will be constant. Detailed balance follows from the presence of a time-reversed trajectory $[q(-t),-p(-t)]$ for any given trajectory. Because the energy is constant along a trajectory, the density $f(l,p)$ at an outflowing boundary will be equal to the corresponding inflowing density $f(l,-p)$ if, and only if, the distribution functions in the two reservoirs are identical functions of energy (i.e., in equilibrium). If we focus upon a differential element of the trajectory, the condi-

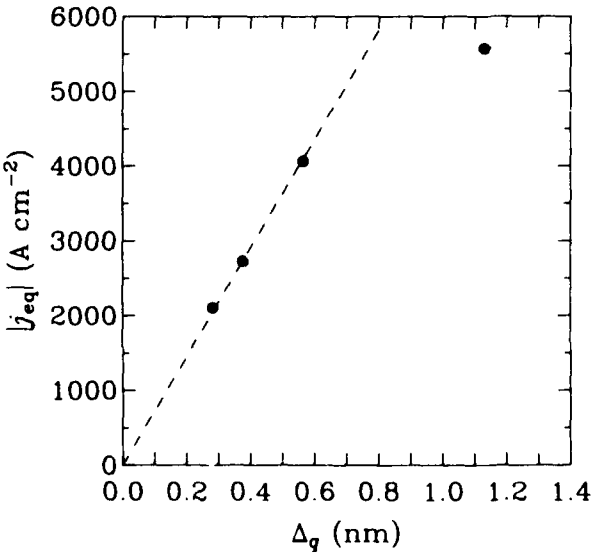


FIG. 23. Violation of the principle of detailed balance in the discrete open-system model. The current density calculated for an asymmetric structure in equilibrium is plotted versus the mesh spacing used in the calculation. The results show that the current density (which measures the departure from detailed balance) is of $O(\Delta_q)$ and is thus a result of the discretization, not of the open-system boundary conditions.

tion that there exists a time-reversed trajectory can be expressed as

$$(\mathcal{L}/i\hbar)(q_1,p_1;q_2,p_2)\stackrel{?}{=}(\mathcal{L}/i\hbar)(q_2,-p_2;q_1,-p_1)\ , \tag{7.12}$$

which becomes $\mathcal{L}=\mathcal{L}^\dagger$ when transformed back to the density-matrix representation, leading to the unsurprising conclusion that time reversibility is equivalent to the Hermiticity of $\mathcal{L}$. In fact, the irreversible model (4.19) satisfies condition (7.12) if we include the boundary terms (4.20). It would appear appropriate to include these terms in the detailed-balance test, whereas we neglect them in the stability analysis. However, this argument leads to the conclusion that the model ought to satisfy detailed balance exactly.

A further consideration of the classical case suggests that the departure from detailed balance might be traceable to discretization errors in the classical trajectories. That is, when we restrict the distribution function to a discrete mesh of points, a particle cannot exactly follow the proper trajectory, and the time-reversed trajectory might not exactly balance it. The way to correct such a situation is to adopt the Lagrangian coordinates discussed in Appendix C. Then the upwind difference would be applied to the directional derivative along a trajectory and would exactly satisfy time reversibility. However, this does not help in cases such as quantum-mechanical tunneling, in which trajectories cannot be defined. Discretization errors in the trajectories would presumably lead to the conclusion that both Δ_q and Δ_p contribute to the error, contrary to what has been observed. If the error were of the form $O(\Delta_q)+O(\Delta_p)$ and the terms had coefficients of different magnitudes, the numerical experiments might easily have overlooked the weaker dependence.

Another way to view the problem of detailed balance in a completely quantum-mechanical context is to note that the equilibrium distribution should satisfy the Bloch equation (6.30). The stationarity of such a distribution under time evolution by the Liouville operator would follow from $[\mathcal{L},\mathcal{H}_{\cdot,\cdot}]=0$. We have noted that this is necessarily true in a closed system, but it is not true for an open-system model. In the present case the commutator has nonzero elements adjacent to the boundaries of the system. These might be removed by including the inhomogeneous terms, but the meaning of an inhomogeneous term in a commutator is far from clear.

The connection between detailed balance and reversibility or Hermiticity suggests the following conjecture: that it is impossible to satisfy exactly both detailed balance and the stability condition (irreversibility) in a model with a finite number of degrees of freedom (such as a bounded, discrete model). That this is possible in a model with an infinite number of degrees of freedom, as in unbounded or continuous models, is the thrust of the conventional theories of irreversibility. If this conjecture is correct, this is a significant limit on the accuracy achievable with discrete open-system models.

D. Comparison of discrete models

Table I summarizes the results of this analysis of the discrete open-system model. It also contains results for other discretization schemes that have been used for similar calculations. The schemes included in the table are, first, the present upwind-difference approximation to the $\mathcal{T}$ operator, denoted "Upwind." Second is the centered-difference approximation studied by Jensen and Buot (1989a) to resolve the problem of detailed balance. In this approximation the kinetic-energy superoperator $\mathcal{T}^{\rm ctr)}$ (for centered difference) becomes

$$\mathcal{T}^{\rm ctr)}_{jk;j'k'}=-\frac{p_k}{2m\Delta_q}\delta_{kk'}(\delta_{j+1,j'}-\delta_{j-1,j'})\ . \tag{7.13}$$

The third column presents an analysis of a centered-difference approximation with upwind differencing applied only at the outflowing boundaries. This is the $\Delta_t\to 0$ limit of the Lax-Wendroff discretization (with upwind differencing at the boundaries) used by Kluksdahl, Kriman, Ferry, and Ringhofer (1989). The continuous time limit is invoked in the present analysis to remove any artifacts of time discretization and thus evaluate this scheme on the same basis as the others. This yields the superoperator $\mathcal{T}^{\rm cub)}$ (for centered, upwind boundary):

TABLE I. Order of errors in discrete open-system models.

$\mathcal{T}$ discretization	Definition	Reference	Continuity	Momentum balance	Detailed balance	Stability: max Imλ
Upwind	Eq. (4.19)	a	$\varepsilon_0[V]$	$O(\Delta_q)+\varepsilon_1[V]$	$O(\Delta_q)$	$-O(\Delta_p)$
Centered	Eq. (7.13)	b	$\varepsilon_0[V]$	$\varepsilon_1[V]$		$+O(\Delta_q^{\ 3})$
Centered upwind boundaries	Eq. (7.14)	c	$O(\Delta_q)+\varepsilon_0[V]$	$O(\Delta_q)+\varepsilon_1[V]$		$-O(\Delta_q\Delta_p)$
Density matrix	Eqs. (3.10),(3.14)	d	0	0	0	$+O(\Delta_t^{\ 1})$

[a] Frensley, 1987a.
[b] Jensen and Buot, 1989a.
[c] Kluksdahl *et al.*, 1989.
[d] Frensley, 1985.

$$\mathcal{T}^{(\mathrm{cub})}_{jk;j'k'} = -\frac{p_k}{m\Delta_q}\delta_{kk'}\times\begin{cases}\frac{1}{2}\delta_{j+1,j'}-\frac{1}{2}\delta_{j-1,j'} & \text{for } p_k<0 \text{ and } j>1\\ \delta_{j+1,j'}-\delta_{j,j'} & \text{for } p_k<0 \text{ and } j=1\\ \frac{1}{2}\delta_{j+1,j'}-\frac{1}{2}\delta_{j-1,j'} & \text{for } p_k>0 \text{ and } j<N_q\\ \delta_{j,j'}-\delta_{j-1,j'} & \text{for } p_k>0 \text{ and } j=N_q\end{cases}. \tag{7.14}$$

The last column of Table I summarizes the time-reversible model based upon the density matrix that we explored in Sec. III.

The errors in the continuity and momentum balance relations were determined by analysis of the discrete equations in the manner described above. These errors include contributions from the potential superoperator $\mathcal{V}$ as well as from the kinetic-energy superoperator $\mathcal{T}$. Because the different discretization schemes that can be used for $\mathcal{V}$ are independent of those for $\mathcal{T}$, the error contributions from $\mathcal{V}$ (denoted as $\varepsilon_i[\mathcal{V}]$) discussed in Sec. VII.B are tabulated separately in Table II. The density-matrix model of Sec. III is set up so as to exactly satisfy the continuity and momentum balance equations. This is possible because the $\mathcal{V}_{(-)}$ superoperator can be evaluated in closed form when applied to the density matrix in real space, but must be approximated by Eq. (4.14), (7.10), or some similar expression when applied to a Wigner function.

The centered-difference form (7.13) also exactly satisfies the continuity and momentum balance equations, if we associate the current density J with the mesh points rather than with the intervals as in Eq. (7.2). In the case of the centered, upwind boundary scheme (7.14), the change in the discretization of the gradient necessarily introduces errors of $O(\Delta_q)$ into all the moment equations. It can be argued that such errors are in some way less significant because the occur only adjacent to the boundaries, but a central lesson of the present analysis is that the boundary terms affect the entire solution, and their influence is not localized to the regions near the boundaries.

The considerations that bear upon departures from detailed balance have been discussed above. The approach described, studying the scaling properties of the equilibrium solutions to the Liouville equation as illustrated in Fig. 23, does not work for the centered-difference (7.13) or centered-upwind boundary (7.14) discretizations because one cannot directly solve for the steady-state distributions with these schemes. Both of them possess at least one spurious mode whose eigenvalue is very close to zero, which in regions of constant potential is of the form $\cos\pi j\Delta_q$, so that its sign alternates between adjacent mesh points. If one attempts to solve for the steady-state distribution, a relatively arbitrary fraction of this mode is incorporated into the solution, rendering the results meaningless. Nevertheless, the considerations previously discussed strongly suggest that the discretization [Eq. (7.14)], at least, probably violates detailed balance to the same order as the upwind-difference scheme. The status of the centered-difference scheme (7.13) is more problematical. Jensen and Buot (1989a) obtained improved results in the sense of a small equilibrium current with this scheme, but it does not seem to be particularly distinguishable from the others on the basis of the symmetry property (7.12) or its commutator with $\mathcal{H}_{(+)}$. The density-matrix approach is presumed to satisfy detailed balance exactly because it is time reversible.

The stability condition is, of course, absolutely essential for a useful model. It is expressed in Table I by the scaling order of the greatest imaginary part of an eigenvalue. The scaling properties of the different discretizations were investigated by a procedure similar to that illustrated in Fig. 23. Both the upwind-difference and centered-upwind boundary schemes are stable, as we expect (all imaginary parts are negative), but the scaling is different. This is illustrated in Fig. 24, which shows the eigenvalue spectrum for the centered-upwind boundary scheme for the same structure used previously. While all the eigenvalues lie in the lower half-plane, they are clustered much nearer the real axis than those of the upwind scheme illustrated in Fig. 9. The centered-difference and the density-matrix schemes are not stable, as they possess eigenvalues with positive imaginary parts. (It should be noted that the specific results obtained for the centered-difference scheme are somewhat suspect. The Δ_q^{-5} dependence is suspiciously close to that of the total number of arithmetic operations required to diagonalize the operator, Δ_q^{-6}, so there is a strong possibility that what was observed here is just the cumulative effect of roundoff errors.)

In summary, no model exactly satisfies all the conditions one would desire. One must therefore decide which model to use on the basis of what is most important for a given application. The information in Tables I and II provides the basis for making such a decision. The analyses that are summarized in the tables, while somewhat tedious, will be useful at two different levels. The first is as a summary of the properties of the different discretiza-

TABLE II. Error terms due to discretization of the potential: $\varepsilon_i[\mathcal{V}]$.

$\mathcal{V}$ discretization	Reference	Continuity ($\varepsilon_0[\mathcal{V}]$)	Momentum balance ($\varepsilon_1[\mathcal{V}]$)
Eq. (4.14)	a	0	$O(1)$
Eq. (7.10)	b	$O(\Delta_p)$	0

[a]Frensley, 1987a; Kluksdahl *et al.*, 1989.
[b]Mains and Haddad, 1988c.

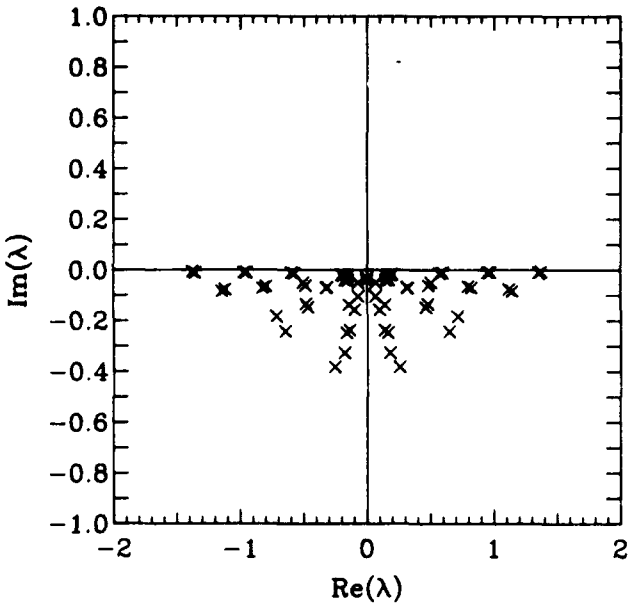


FIG. 24. Eigenvalue spectrum resulting from the discretization (7.14). This discretization results in a stable model.

tion schemes studied here. At a more general level the present analyses provide an example of the sort of study required to make sense of the multitude of discretization schemes for a given physical problem.

VIII. CONCLUSIONS

The central conclusion of the present work is that an open system, in the sense of one that exchanges particles with its environment through spatially localizable interfaces, is necessarily irreversible. The reasoning behind this conclusion is a *reductio ad absurdum* argument. We have seen that a particular reversible model of an open system possesses unphysical instabilities. The mathematical properties underlying these instabilities, namely the existence of complex eigenvalues of non-Hermitian superoperators and the requirement that these occur in conjugate pairs due to time-reversal symmetry, are sufficiently general that we should expect such instabilities in any reversible model. Thus physically acceptable models of open systems must be inherently time irreversible.

A particular class of irreversible open-system models was presented, and the stability of the resulting solutions was demonstrated. The irreversibility of these models follows from making a distinction between particles entering and leaving the system. Similar ideas, generally applied in the time domain, are the basis for the established theories of irreversibility and dissipation. The present work demonstrates that spatial boundary conditions can be used to introduce irreversibility in a way very similar to that by which temporal initial conditions do so.

The present study of the kinetic theory of open systems helps to clarify the roles of superoperators generated by the commutator and anticommutator of a physical observable. It was demonstrated that, at the kinetic level, only the commutator superoperators should acquire non-Hermitian parts to model irreversible phenomena. Anticommutator superoperators remain Hermitian and are used to evaluate expectation values.

Some of the more mathematical issues concerning the properties of the present open-system models remain unresolved, particularly the question of positivity of the resulting Wigner distribution functions. However, the results obtained by applying these models to the resonant-tunneling diode demonstrate the usefulness and credibility of this approach.

This work is certainly not an exhaustive examination of the theory of open systems. Undoubtedly, many more approaches to the subject can be formulated. However, one should note that the significant behaviors of an open system involve a strong coupling between the system and its environment and large deviations from equilibrium within the system. It thus appears unlikely that perturbative approaches will contribute much to the theory of such systems. Other analytic approaches will be effective only in cases displaying some exceptional symmetry (and of course the present definition of open system rules out translational symmetry). It thus appears that numerical models such as those examined here will probably be the mainstay of such investigations.

Note added in proof: Three recent results in this field have come to the author's attention: Jensen and Buot [J. Appl. Phys. **67**, 2153 (1990)] have studied a second-order differencing scheme for evaluation of the Wigner function, and they find that this improves the results for the resonant-tunneling diode in several respects. Govindan, Grubin, and de Jong have reported an open-system boundary condition for the density matrix (in real space) which appears to avoid the instabilities discussed in Sec. III. The boundary condition involves the specification of both the density and the current. Register, Ravaioli, and Hess have developed an improved traveling-wave boundary-condition scheme for the time-dependent Schrödinger equation. The latter two works will appear in the *Proceedings of the Workshop on Computational Electronics*, University of Illinois–Urbana, May 21–22, 1990, edited by K. Hess, J.-P. Leburton, and U. Ravaioli (Cluyer, Norwell, MA, in press).

ACKNOWLEDGMENTS

The author would like to acknowledge helpful discussions with J. R. Barker, R. T. Bate, C. D. Cantrell, D. K. Ferry, K. Hess, K. L. Jensen, N. C. Kluksdahl, A. J. Leggett, R. Lodenkamper, J. H. Luscombe, R. K. Mains, F. J. Narcowich, W. Pötz, M. A. Reed, and L. E. Reichl. This work was supported in part by the Office of Naval Research, the Defense Advanced Research Projects Agency, and the U.S. Army Research Office.

APPENDIX A: SELF-CONSISTENT POTENTIAL OF A TUNNELING STRUCTURE

The semiconductor heterostructure used in the analysis in Sec. I.B consisted of an undoped 3.39 nm (12 unit cells) layer of $Al_{0.3}Ga_{0.7}As$ embedded in GaAs crystal doped such that the mobile-electron density was 6×10^{17} cm^{-3}, and the temperature was 300 K. (This particular structure was chosen to provide a clear demonstration of the failure of the standard tunneling theory.) The calculations were done for a bias of 0.2 V ($\approx 8kT$) applied to the structure.

The initial approximation for the self-consistent potential was obtained from a generalized (to finite temperature) Thomas-Fermi screening approximation. At its most fundamental level, the Thomas-Fermi approximation can be viewed as an expression for the Wigner distribution function:

$$f(\mathbf{x},\mathbf{k})\approx\frac{1}{1+e^{\beta[T(\mathbf{x},\mathbf{k})+v(\mathbf{x})-\mu]}}\ , \tag{A1}$$

where $T(\mathbf{x},\mathbf{k})$ is obtained by taking the kinetic-energy term of the Hamiltonian in the neighborhood of $\mathbf{x}$, extending this form over all space, and taking the expectation value of the resulting operator on the plane-wave state $|\mathbf{k}\rangle$. This typically gives $T(\mathbf{x},\mathbf{k})=\hbar^2k^2/2m^*(\mathbf{x})$, where the effective mass m^* can vary with position, as discussed in Appendix E. Integrating over all momenta gives the more familiar expression (Blakemore, 1982)

$$n(\mathbf{x})=N_c\mathcal{F}_{1/2}\{\beta[\mu-v(\mathbf{x})]\}\ , \tag{A2}$$

where $N_c=2(m^*/2\pi\hbar^2\beta)^{3/2}$ is the "effective density of states," and $\mathcal{F}_{1/2}$ is the Fermi-Dirac integral of order $\frac{1}{2}$. The potential v can be separated into a Hartree potential v_H and a "heterostructure" potential v_s which describes the heterostructure band offsets:

$$v(\mathbf{x})=v_H(\mathbf{x})=v_s(\mathbf{x})\ . \tag{A3}$$

The Hartree potential satisfies Poisson's equation,

$$-\nabla\cdot\epsilon\nabla v_H=e^2[n(\mathbf{x})-N_d(\mathbf{x})]\ , \tag{A4}$$

where N_d is the background positive charge density (ionized donor density). Inserting Eqs. (A2) and (A3) into (A4) produces a Poisson equation with a nonlinear source term, which is readily solved in a finite-difference approximation by a multidimensional Newton iteration technique (Selberherr, 1984, Chap. 7). The boundary conditions for Eq. (A4) are obtained from the requirement that the system asymptotically approach charge neutrality,

$$v_H=\mu-v_s-\frac{1}{\beta}\mathcal{F}_{1/2}^{-1}\left[\frac{N_d}{N_c}\right]\ , \tag{A5}$$

with all quantities evaluated in the appropriate asymptotic region. In practice, these boundary conditions are applied at fixed locations sufficiently distant that charge neutrality is well satisfied (see Fig. 1). Note that the reference energy for v_s may be chosen arbitrarily; this reference and the externally imposed μ then uniquely determine v_H. Strictly speaking the Thomas-Fermi approximation is only an equilibrium approximation. However, in some structures, such as the present single-barrier device, one can identify regions in which a local quasiequilibrium ought to hold. In such cases one can obtain useful results for the nonequilibrium case by assuming that the chemical potentials differ from one region to another, as illustrated in Fig. 1.

To evaluate the self-consistent potential within the conventional independent-electron tunneling theory, we need to define precisely the (mixed) quantum state of the system. The fundamental assumption of tunneling theory is that the electrons will be found in the eigenstates of the Hamiltonian (generally un-normalizable scattering states), and the probability of occupation of the left- and right-incident states is given by the different Fermi distributions of the respective contacts. We may summarize these assumptions by writing a density operator for the system

$$\begin{aligned}\rho(x,x')=&\int_{v_l}^{\infty}\frac{dE}{2\pi\hbar s_l(E)}f_\perp(E-\mu_l)\psi_l(E,x)\psi_l^*(E,x')\\&+\int_{v_r}^{\infty}\frac{dE}{2\pi\hbar s_r(E)}f_\perp(E-\mu_r)\psi_r(E,x)\\&\times\psi_r^*(E,x')\ ,\end{aligned} \tag{A6}$$

where $v_{l,r}$ are the asymptotic potentials to the left and right, and $s_{l,r}(E)$ is the velocity of an electron of energy E at the respective boundary. Here $f_\perp$ is the Fermi-Dirac distribution function integrated over the transverse momenta:

$$f_\perp(E)=(m^*/\pi\hbar^2\beta)\ln(1+e^{-\beta E})\ . \tag{A7}$$

The $\psi_{l,r}$ are the solutions of Schrödinger's equation in an effective-mass approximation,

$$-\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*(x)}\frac{\partial}{\partial x}\psi+v\psi=E\psi\ , \tag{A8}$$

with unit incident amplitude from the left or right, respectively. Using Eq. (A6) we can evaluate any physical observable of the tunneling system, although, in the literature, the content of (A6) is usually expressed only in an equation for the current density. However, to evaluate the self-consistent potential we need to evaluate the electron density, which is simply $n(x)=\rho(x,x)$. Inserting this into Poisson's equation (A4) and again applying the condition (A5) at each boundary, we obtain the potential shown by the dashed line in Fig. 1. This potential is clearly unphysical, as discussed in the text, because inelastic processes are neglected. A proper description of such processes requires a kinetic theory.

The quantum-kinetic calculations shown in Fig. 3 were performed by solving the steady-state kinetic equation (4.27) and Poisson's equation (A4) self-consistently, again by a multidimensional Newton iteration scheme. The electron density n in Poisson's equation was obtained from the Wigner function using Eq. (7.1). Phonon

scattering was included by adding the Boltzmann collision operator described in Appendix F, for both longitudinal-optic and acoustic phonons, to the Liouville operator used in Eq. (4.27). The calculation of Fig. 3(a) assumed fixed boundary distributions [Eq. (5.1)]. The calculation of Fig. 3(b) assumed displaced equilibrium boundary distributions, to take into account the transport processes in the contacting layers (Mains and Haddad, 1988c). These distributions were just Eq. (5.1) with argument $p_k - p_0$, where $p_0 = -\mu_e m \partial v/\partial x$, μ_e is the electron mobility (taken to be 5000 $\mathrm{cm^2\,V^{-1}\,s^{-1}}$), and the electric field was evaluated at the respective boundary. This shifts the distribution function so that a greater density of electrons enters on the upstream side and a lesser density enters on the downstream side, which makes the screening of the electric field more effective.

Other self-consistent calculations of far-from-equilibrium tunneling structures have focused upon the double-barrier resonant-tunneling diode because of its greater technological significance. Cahay *et al.* (1987) performed a self-consistent Schrödinger calculation, as described above. However, they assumed a device structure with undoped spacer layers on either side of the double barrier. The contact potentials of the doped-undoped junctions created an additional energy barrier which, by confining the electrons, helped to enforce charge neutrality, and thus the unphysical effects described above were avoided. If the undoped spacer layers had not been present, an unphysical potential would have been obtained.

Pötz (1989) also performed a self-consistent Schrödinger calculation. In this case the unphysical results were avoided by modifying the definition of the electron ensemble from (A6) to one in which the notch states were weighted with the Fermi distribution of the upstream electrode, in effect assuming a high rate of inelastic processes to fill these states. A displaced distribution function as described above was also used in this calculation, but the drift momentum was chosen so as to satisfy charge neutrality, rather than to approximate ohmic conduction.

Kluksdahl *et al.* (1989) performed a self-consistent kinetic (Wigner-function) calculation of the type described above, with a relaxation-time approximation for the collision operator. The results showed an unphysically large electric field at the upstream boundary. Similar results were obtained by the present author (Frensley, 1989a, 1989b) from a kinetic model lacking the collision term. As in the single-barrier case, the inclusion of phonon collisions and displaced boundary distributions led to more credible results (that is, more complete screening of the field) for the self-consistent potential (Mains and Haddad, 1988c; Frensley 1989a, 1989b).

APPENDIX B: VIOLATION OF CONTINUITY IN THE PAULI MASTER EQUATION

The Pauli master equation (see Kreuzer, 1981, Chap. 10) is derived under the assumption that the density matrix is and remains diagonal in the basis of eigenstates of the Hamiltonian,

$$\rho(x,x';t) = \sum_i P_i(t)\psi_i(x)\psi_i^*(x') , \tag{B1}$$

where $P_i(t)$ is the probability of the system to be in state i. The master equation is then

$$dP_i/dt = \sum_j [W_{ij}P_j(t) - W_{ji}P_i(t)] , \tag{B2}$$

where the W_{ij} are the golden-rule transition rates. Consider transitions from a state i to a state j which have different spatial distributions: $|\psi_i(x)|^2 \neq |\psi_j(x)|^2$. Then the rate of change of the density is

$$\begin{aligned}\frac{\partial}{\partial t}\rho(x,x;t) &= \frac{\partial P_i}{\partial t}|\psi_i(x)|^2 + \frac{\partial P_j}{\partial t}|\psi_j(x)|^2 \\ &= [W_{ji}P_i(t) - W_{ij}P_j(t)] \\ &\quad \times [|\psi_j(x)|^2 - |\psi_i(x)|^2] .\end{aligned} \tag{B3}$$

However, i and j are eigenstates of the Hamiltonian, which means that $\langle i|\mathbf{J}|i\rangle$ and $\langle j|\mathbf{J}|j\rangle$ are constant (for scattering states) or even zero (for bound states). In either case,

$$\nabla\cdot\langle i|\mathbf{J}|i\rangle = \nabla\cdot\langle j|\mathbf{J}|j\rangle = 0 . \tag{B4}$$

Now, the rate of change of the density will be zero if either of the two bracketed terms in Eq. (B3) is zero. In thermal equilibrium the first term is zero by the principle of detailed balance, but away from equilibrium it is, in general, nonzero. The second term will be zero if the probability distributions of the eigenstates i and j are identical. This happens in only a very few cases, most notably for the plane-wave states of a free particle.

Thus the assumption that the density matrix has the form (B1) for far-from-equilibrium systems will lead, in general, to a violation of the continuity equation.

APPENDIX C: BOUNDARY CONDITIONS FOR LAGRANGIAN-VARIABLE APPROACHES

Broadly speaking, there are two ways to set up a transport problem: the Eulerian approach, in which the coordinates are fixed in the reference frame of the observer; and the Lagrangian approach, in which the coordinates are fixed in the reference frame of the transported fluid. The present work focuses upon the Eulerian approach. However, a number of formulations of quantum-transport theory are expressed in terms of Lagrangian variables. These include the center-of-mass approach of Lei and Ting (1985) and the quantum Langevin-equation approach of Hu and O'Connell (1987). The accelerated basis states studied by Krieger and Iafrate (1986) adapt the Lagrangian variables to pure-state quantum mechanics. It appears that none of these approaches has yet been applied to an open-system problem in the present sense, so there has been no analysis of the effects of

boundary conditions within the Lagrangian approaches. Moreover, it is not at all clear that such approaches are well adapted to the description of tunneling, where there is no classical trajectory (although in this connection one should note the work of Jensen and Buot, 1989b, in which the trajectories in a resonant-tunneling diode were inferred from a solution for the Wigner function).

In the classical case, however, much of the work dealing with open systems (and most of the work treating electron transport in nonuniform systems) has been cast in terms of the Lagrangian variables. This includes both deterministic approaches, such as that of Baranger and Wilkins (1987), and stochastic approaches, such as the widely used Monte Carlo technique (Jacoboni and Reggiani, 1983; Castagné, 1985; Constant, 1985; Reggiani, 1985). If we consider the boundary conditions in such approaches, it becomes apparent that the "inflowing" boundary conditions [Eq. (4.7)] will occur quite naturally. In the approach of Baranger and Wilkins the Lagrangian variables define the mean trajectories of the particles, so one must specify the initial conditions on the trajectory, which is the value of the distribution function at the point where the trajectory enters the domain. Thus the boundary conditions are completely equivalent to Eq. (4.7).

In the case of the Monte Carlo technique the boundary conditions are determined implicitly by the details of the algorithm used in the calculation, and such details are often omitted from the published reports. To understand the relationship between the algorithm and the boundary conditions, let us consider the algorithms described by Lebwohl and Price (1971) and Hockney and Eastwood (1981) (which is also described by Castagné, 1985). Any electron leaving the domain of the Lebwohl and Price calculation is immediately replaced by another electron entering randomly from either contact, with an initial momentum chosen from a thermal distribution. Thus the number of electrons in the system is fixed (and the fact that this leads to simpler and more efficient programs is the motivation for the Lebwohl and Price approach). A distribution function evaluated with this algorithm will satisfy boundary conditions of the form (4.7), but the values of the boundary distributions will not necessarily remain fixed, as they depend upon the rate at which electrons impinge upon the contact. To view the problem another way, the same algorithm would be obtained from a model in which the system was assumed to be periodic, but which had a very strong scattering process located at that plane where the system closed upon itself. Thus this approach really describes a closed system, and the fixed number of particles within the system is an indication that the system is actually closed. A truly open system results if the particles entering the domain are chosen by an independent stochastic process, and the resulting distribution function would then satisfy Eq. (4.7) with fixed boundary distributions. The algorithm described by Hockney and Eastwood (1981) is almost of this form, though the rate of particle injection is adjusted in response to the nearby density.

The discussion of Monte Carlo algorithms and boundary conditions brings out an important point: The number of particles in an open system necessarily fluctuates. While I have not addressed fluctuation phenomena in the present work, a more complete description should deal with such effects.

APPENDIX D: BOUNDARY CONDITIONS FOR SCHRÖDINGER'S EQUATION

The application of Schrödinger's equation to an open system in the present sense is a large part of the formal theory of scattering. The traditional approach is to expand the wave function in a set of traveling waves, at least in the asymptotic region. This implicitly sets the boundary conditions employed in the analysis. With the present interest in the quantum-transport properties of (often complex) fabricated structures, purely numerical techniques for solving Schrödinger's equation have become more important. In these techniques one has a direct representation of the wave function as a complex-valued function of position, typically on a discrete basis (using finite-difference or finite-element techniques, for example). In this situation, the appropriate boundary conditions must be explicitly specified, and the proper choice of boundary conditions is a prerequisite to obtaining any meaningful results.

Let us first consider the steady-state case in a one-dimensional system extending over the interval $0 \le x \le l$. In general, we seek wave functions corresponding to traveling waves incident from either the left or the right. These states will include a reflected component, which appears at the same boundary as the incident wave, and a transmitted component, which appears at the opposite boundary. For example, for an eigenstate incident from the left, we have

$$\psi(x) = Ae^{ik_0 x} + Be^{-ik_0 x} \quad \text{for } x \le 0 , \tag{D1}$$

$$\psi(x) = Ce^{ik_l x} \quad \text{for } x \ge l . \tag{D2}$$

We know the value of A (typically $A = 1$), but we do not know the value of B or C. A straightforward way to evaluate ψ is temporarily to assume $C = 1$, from which we obtain the initial conditions $\psi(l) = 1$ and $\partial\psi(l)/\partial x = ik_l$. The steady-state Schrödinger equation may then be integrated from $x = l$ to $x = 0$, and the solution may then be normalized so that $A = 1$.

A more elegant approach is the quantum transmitting boundary method (QTBM) of Lent and Kirkner (1990). The essence of this approach is to apply *mixed* boundary conditions at each boundary. The mixed boundary conditions involve fixing the value of a linear combination of the wave function and its gradient. At the left-hand boundary,

$$\psi(0) = A + B , \tag{D3}$$

$$\psi'(0) = \partial\psi/\partial x|_0 = ik_0(A - B) . \tag{D4}$$

Solving for A we obtain

$$A=\tfrac{1}{2}[\psi(0)-i\psi'(0)/k_0]\ . \tag{D5}$$

A similar expression for the incident amplitude at the right-hand boundary (let us call it D) may be readily derived:

$$D=\tfrac{1}{2}[\psi(l)+i\psi'(l)/k_l]\ . \tag{D6}$$

Equations (D5) and (D6) are the QTBM boundary conditions. They define an implicit relationship between ψ and ψ' and thus they must be solved along with Schrödinger's equation itself. This is readily done in a numerical approach in which the Schrödinger equation is approximated by a set of algebraic equations: One simply adds (D5) and (D6) to the set and solves them simultaneously. The QTBM is readily extended to two-dimensional problems (Lent and Kirkner, 1990) and to problems involving complex energy-band structures that require more than one basis function per unit cell (Frensley and Luscombe, 1990). Note that the QTBM boundary conditions are energy dependent, this dependence being implicit in the dependence of Eqs. (D5) and (D6) on k_0 and k_l.

If the problem is time dependent (typically because the potential varies with time), the problem of boundary conditions is much more complex. If we start with the knowledge that the electron in question is in a particular eigenstate of the Hamiltonian $H(0)$ at $t=0$, at some later time t when the potential has changed perceptibly the electron will not in general be in an eigenstate of $H(t)$, but will be in a superposition of such eigenstates. Let us focus our attention on the boundary at $x=0$ and assume that the potential does not vary in its immediate neighborhood. The wave function with unit incident amplitude will be of the general form

$$\psi(x,t)=e^{i(kx-\omega t)}+\phi(x,t)\ , \tag{D7}$$

and all we know about the reflected wave $\phi(x,t)$ is that it is a solution of Schrödinger's equation and all of its momenta should be negative. (However, a momentum-space expansion of ϕ is not feasible because we wish to deal only with ϕ over a small range in x.) Mains and Haddad (1988a) have reported calculations of the transient response of a resonant-tunneling diode using

$$\phi(x,t)=B(x,t)e^{i(-kx-\omega t)}\ , \tag{D8}$$

with $B(x,t)$ assumed to be slowly varying in space and time. Inserting Eq. (D8) into Schrödinger's equation gives

$$\frac{\partial B}{\partial t}=\frac{\hbar k}{m}\frac{\partial B}{\partial x}+\frac{i\hbar}{2m}\frac{\partial^2 B}{\partial x^2}\ . \tag{D9}$$

Mains and Haddad used the first-derivative term of Eq. (D9) to update the value of $B(0,t)$ (Dirichlet boundary condition) in a time-integration procedure. This amounts to looking a short distance into the domain to determine what is coming out.

Let us consider another scheme for determining the boundary condition, which I have not tested in a practical computation, but which has the pedagogical advantage of explicitly displaying the non-Markovian nature of the problem. Suppose that we are implementing a discrete time-integration scheme with step size Δ_t and that we wish to apply a Neumann spatial boundary condition at $x=0$. Then we need a way to determine the value of $\partial\phi/\partial x$ at the next time step. We Fourier transform Schrödinger's equation and solve for k to obtain

$$ik=\pm i\sqrt{2m(\hbar\omega-v)}/\hbar\ . \tag{D10}$$

In the case of the reflected waves propagating out of the $x=0$ boundary we would choose the negative sign on the square root. Now suppose that we approximate the right-hand side of Eq. (D10), over an appropriate range of energies, by a polynomial in $-i\omega$:

$$ik\approx\sum_{n=0}^{N}a_n(-i\omega)^n\ .$$

Inverting the Fourier transform, we obtain an expression for the gradient of ϕ:

$$\frac{\partial\phi}{\partial x}(0,t_0)\approx\sum_{n=0}^{N}a_n\frac{\partial^n\phi}{\partial t^n}(0,t_0)\approx\sum_{m=0}^{N}b_m\phi(0,t_0-m\Delta_t)\ , \tag{D11}$$

where the latter expression is a finite-difference approximation to the differential operator, and we approximate this operator using only the values of ϕ at times prior to t_0 because those are the only known values. (Thus the time-reversal symmetry is broken.) Note that Eq. (D11) explicitly demonstrates the dependence of the boundary condition on the prior history of the system and thus shows its non-Markovian character. The finite-difference coefficients b_m may be obtained from the a_n by expanding $\phi(0,t_0-m\Delta_t)$ in a Taylor series. One thus obtains the set of equations

$$a_n=\sum_{m=0}^{N}\frac{(-m\Delta_t)^n}{n!}b_m\ , \tag{D12}$$

which must be solved to find the b_m.

The essence of this scheme is that we use the previously calculated values of the wave function at the boundary to attempt to predict the next value of the gradient. This is a particular example of linear prediction (Makhoul, 1975). It also illustrates a general property of derivations of irreversible phenomena in quantum mechanics: When one attempts to remove (or at least ignore) the effects of some of the degrees of freedom in a system (in this case the spatial locations outside the boundary), they reassert themselves in the time domain, in the form of non-Markovian terms (Zwanzig, 1964).

APPENDIX E: POSITION-DEPENDENT EFFECTIVE MASS

In the semiconductor structures that originally motivated this work the charge carriers whose motion we

seek to describe are really quasiparticles, whose properties are determined by the energy-band structure (or energy-momentum dispersion relation) of the semiconductor material. These carriers usually occupy states near an extremum of a band, and thus for the simpler cases of interest the band structure can be approximated as

$$E(k) \approx v_s + (\hbar^2/2m^*)(k - k_0)^2 , \tag{E1}$$

where v_s is the energy at the edge of the band and is just the heterostructure potential used in Appendix A, k_0 is the wave vector at which this extremum occurs, and m^* is the "effective mass" that characterizes the curvature of the dispersion relation. This dispersion relation may be modeled by the effective-mass Schrödinger equation

$$i\hbar \partial\Psi/\partial t = -(\hbar^2/2m^*)\nabla^2\Psi + (v_s + v_H)\Psi , \tag{E2}$$

where v_H is the Hartree potential, which is assumed to be slowly varying. The wave function Ψ in Eq. (E2) is strictly an envelope function for the true wave function. In the Wannier-Slater approach to effective-mass theory (Slater, 1949), Ψ is a discrete function (defined on the lattice points) giving the amplitude of the Wannier function at each point [though Ψ is approximated by a continuous function to derive the differential equation (E2)]. In the approach of Luttinger and Kohn (1955), Ψ is a continuum but band-limited function, which is multiplied by a perfectly periodic Bloch function to obtain the complete wave function.

A semiconductor heterostructure is a single crystal that includes (deliberately introduced) local changes in the chemical composition. These introduce changes in the "local band structure" which must be incorporated into the effective-mass equation (E2) to obtain an accurate model of the quasiparticle dynamics in a heterostructure. For the sake of concreteness let us consider an abrupt heterojunction. The local band-edge energy v_s will be shifted across the heterojunction, and this effect is easily incorporated into Eq. (E2) by making v_s a function of position. In general, the value of the effective mass will also change across a heterojunction, and this requires a more careful treatment of the kinetic-energy term. (Another way to view this problem is to state the conditions for matching Ψ across an interface with discontinuous m^*. Because the matching condition follows uniquely from the form of the Hamiltonian, we shall focus upon the latter.) The problem is that many of the expressions one might write down [such as that in (E2)] become non-Hermitian when m^* is taken to be a function of position. The simplest manifestly Hermitian form is

$$T = -\frac{\hbar^2}{2}\frac{\partial}{\partial x}\frac{1}{m^*(x)}\frac{\partial}{\partial x} , \tag{E3}$$

although other, more complicated expressions have been suggested (see Morrow and Brownstein, 1984). In general, it appears that Eq. (E3), which might be termed the "minimal Hermitian form," is an adequate approximation when the magnitude of the change in m^* is small, as is typically true of equivalent energy bands in closely related materials. When the discontinuity is of a larger magnitude, as when inequivalent bands are involved, one probably needs to solve the multiband problem explicitly and infer the form of the effective-mass equation from the results (see, for example, Grinberg and Luryi, 1989).

We can obtain different discrete approximations to (E3) depending upon where we assume the heterojunction to be actually located with respect to the mesh points. The most consistent scheme is to assume that the junction is located midway between two adjacent mesh points. The discrete Hamiltonian (3.9) then becomes (Mains, Mehdi, and Haddad, 1989)

$$\begin{aligned} H_{ii} &= \frac{\hbar^2}{4\Delta_x^2}\left[\frac{1}{m^*_{i-1}} + \frac{2}{m^*_i} + \frac{1}{m^*_{i+1}}\right] + v_i , \\ H_{i,i+1} &= H_{i+1,i} = -\frac{\hbar^2}{4\Delta_x^2}\left[\frac{1}{m^*_i} + \frac{1}{m^*_{i+1}}\right] , \end{aligned} \tag{E4}$$

which was used in all of the tunneling calculations presented here.

If we use Eq. (E3) to construct the kinetic-energy superoperator $\mathcal{T}_{(-)}$, how is the form of this superoperator (in the Wigner-Weyl representation) affected? We might hope that a simple expression would result, such as

$$\mathcal{T}_{(-)} \stackrel{?}{=} \mathcal{P}_{(-)} m^*(q)^{-1} \mathcal{P}_{(+)} . \tag{E5}$$

(This is the expression that was actually used in the calculations presented here.) Unfortunately, Eq. (E5) holds only if

$$m^*(q)^{-1} = \tfrac{1}{2}[m^*(x)^{-1} + m^*(x')^{-1}] ,$$

which holds only if the band structure varies slowly as a function of position. In general, a position-dependent effective mass will produce a nonlocal form for the kinetic-energy superoperator in the Wigner-Weyl representation (Barker, Lowe, and Murray, 1984). A more complete treatment, expressing the Wigner-Weyl transformation in terms of the Wannier and Bloch representations (rather than the position and momentum representations) has been developed by Miller and Neikirk (1990). This analysis also demonstrates a nonlocal kinetic-energy term.

APPENDIX F: THE BOLTZMANN COLLISION SUPEROPERATOR FOR PHONON SCATTERING IN SEMICONDUCTORS

To investigate the full range of phenomena that occur in open systems, one needs a model of the dissipative processes (such as scattering of electrons by phonons in semiconductors) that occur within the system. However, the question of the correct description of such processes is at present far from resolved (see Jauho, 1989). Therefore, in the inductive spirit of the present work, we shall assume *a priori* that the classical Boltzmann collision

operator acting upon the Wigner distribution is an adequate approximation at *some* level. The form that the Boltzmann operator takes within the present one-dimensional model is developed below.

In solid-state physics the name "Boltzmann equation" is applied to any transport equation that combines the Liouville description of ballistic motion with a local Markovian model of the stochastic processes. This can include such processes as the scattering of electrons by phonons or impurities. These will be considered to be one-body processes because the phonon and impurity degrees of freedom are not explicitly included in the model, and thus (neglecting Fermi degeneracy) such processes lead to terms linear in the distribution function. The Boltzmann equation can also include a master-operator description of two-body interactions such as electron-electron scattering (and in statistical physics the name "Boltzmann equation" usually refers more specifically to this kinetic equation), and such a term will be nonlinear in the single-particle distribution function (assuming the Stosszahlansatz). For the present purposes we shall consider only one-body interactions so that the collision operator is linear.

The Boltzmann collision term is usually written in the form (Ferry, 1980)

$$(\mathcal{C}_B f)(q,k)=\int \frac{dk'}{2\pi}[W_{kk'}f(q,k')-W_{k'k}f(q,k)] , \quad \text{(F1)}$$

where $W_{kk'}$ is the rate of scattering from plane-wave state k' to state k. (To maintain consistency with the literature, we shall use the wave vector to label these states, rather than the momentum.) Equation (F1) can be rewritten to emphasize the linear, homogeneous nature of the collision term:

$$\begin{aligned}(\mathcal{C}_B f)(q,k)&=\int \frac{dk'}{2\pi}\left[W_{kk'}-\delta(k-k')\int dk'' W_{k''k}\right]\\&\quad\times f(q,k')\\&\equiv\int \frac{dk'}{2\pi}C_B(k,k')f(q,k') .\end{aligned} \quad \text{(F2)}$$

The collision term is local, so that in the complete kernel of $\mathcal{C}_B$ there is a δ function in q, which is suppressed from the above definition. Note that the potential superoperator $\mathcal{V}$ has a similar dependence on q [Eq. (4.10)], and as a result $\mathcal{C}_B$ and $\mathcal{V}$ have the same sparsity structure in the discrete approximation [see Eq. (4.26)]. Thus the addition of $\mathcal{C}_B$ to the calculation requires no modification to the superoperator data structures or solution procedures.

The scattering rates $W_{kk'}$ are taken to be the Fermi golden-rule rates. For electron-phonon scattering,

$$W_{\mathbf{k}\mathbf{k}'}=\frac{2\pi}{\hbar}|\langle\mathbf{k}|H_{ep}|\mathbf{k}'\rangle|^2\delta(E_{\mathbf{k}}-E_{\mathbf{k}'}\mp\hbar\omega) , \quad \text{(F3)}$$

where H_{ep} is the Hamiltonian for the electron-phonon interaction and ω is the phonon frequency. In Eq. (F3) and the following, the upper sign refers to phonon absorption and the lower sign refers to phonon emission. The transition rates depend upon the full three-dimensional $\mathbf{k}$ of each state, whereas the numerical calculations at present consider only the longitudinal momentum k. Thus the scattering rates must be "projected" onto the one-dimensional model. To do so, we first assume that the distribution of electrons with respect to the transverse momenta of the initial state $\mathbf{k}'_\perp$ is a normalized Maxwellian distribution at a fixed temperature:

$$f(q,\mathbf{k}')=f_\parallel(q,k'_\parallel)f_\perp(q,\mathbf{k}'_\perp) , \quad \text{(F4)}$$

where

$$f_\perp(\mathbf{k}'_\perp)=2\pi\lambda_T^2\exp(-\lambda_T^2\mathbf{k}'^2_\perp/2) , \quad \text{(F5)}$$

with λ_T defined in Eq. (3.3). The resulting scattering rates are then integrated over the transverse momenta of the final states:

$$\begin{aligned}W_{k_\parallel,k'_\parallel}=\frac{\Omega\lambda_T^2}{(2\pi)^2\hbar}\int d^2\mathbf{k}_\perp\int d^2\mathbf{k}'_\perp|\langle\mathbf{k}|H_{ep}|\mathbf{k}'\rangle|^2\\\times\delta(E_{\mathbf{k}}-E_{\mathbf{k}'}\mp\hbar\omega)\\\times\exp(-\lambda_T^2\mathbf{k}'^2_\perp/2) ,\end{aligned} \quad \text{(F6)}$$

where Ω is the volume of the crystal. Henceforth we shall drop the subscript from the $k_\parallel$.

For polar optical-phonon scattering the absolute square of the matrix element is (in SI notation and from Fawcett, Boardman, and Swain, 1970)

$$\begin{aligned}|\langle\mathbf{k}|H_{\rm po}|\mathbf{k}'\rangle|^2=\frac{2\pi\hbar\omega_{\rm lo}}{\Omega|\mathbf{k}-\mathbf{k}'|^2}\left|\frac{e^2}{4\pi\epsilon_0}\right|\\\times\left|\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_{\rm dc}}\right|\left|N_{\rm lo}+\begin{matrix}0\\1\end{matrix}\right| ,\end{aligned} \quad \text{(F7)}$$

where $\omega_{\rm lo}$ is the longitudinal-optical phonon frequency, and $\epsilon_{\rm dc}$ and ϵ_∞ are the low- and high-frequency permittivities of the semiconductor, respectively. The phonon occupation number $N_{\rm lo}$ is given by the Bose-Einstein distribution, and again the upper term (0) refers to absorption and the lower term (1) accounts for spontaneous emission. The one-dimensional scattering rates are obtained by inserting Eq. (F7) into (F6). After some manipulation, one can write an expression for the scattering rate. First, define dimensionless quantities a and b as

$$a=\left|\frac{\lambda_T^2k'(k-k')\mp\beta\hbar\omega_{\rm lo}}{\sqrt{2}\lambda_T(k-k')}\right| ,$$

$$b=\left|\frac{\lambda_T^2k(k-k')\mp\beta\hbar\omega_{\rm lo}}{\sqrt{2}\lambda_T(k-k')}\right| .$$

Then the scattering rate is

$$\begin{aligned}W_{kk'}^{(\rm po)}=2\pi\beta\omega_{\rm lo}\left|\frac{e^2}{4\pi\epsilon_0}\right|\left|\frac{1}{\epsilon_\infty}-\frac{1}{\epsilon_{\rm dc}}\right|\left|N_{\rm lo}+\begin{matrix}0\\1\end{matrix}\right|\\\times\frac{e^{a^2}}{\lambda_T|k-k'|}\left|\frac{\pi}{2}\right|^{1/2}\operatorname{erfc}[\sup(a,b)] .\end{aligned} \quad \text{(F8)}$$

The collision operator for polar optical-phonon scattering in the one-dimensional model is then obtained by inserting Eq. (F8), for both phonon emission and absorption, into a discretized version of (F2).

The collision operator for acoustic deformation-potential scattering may be similarly constructed. Assuming equipartition of energy in the acoustic modes, the matrix element is (Fawcett, Boardman, and Swain, 1970)

$$\langle \mathbf{k}|H_{\rm ap}|\mathbf{k}'\rangle|^2=\frac{\hbar\Xi_a^2}{2\rho_m s\Omega}|\mathbf{k}-\mathbf{k}'|N_\omega\approx\frac{\Xi_a^2}{2\beta\rho_m s^2\Omega}\ , \qquad \text{(F9)}$$

where Ξ_a is the acoustic deformation potential, ρ_m is the mass density of the material, and s is the velocity of sound. The second expression is obtained by expanding the Bose distribution for low energies using $\omega=s|\mathbf{k}-\mathbf{k}'|$. Inserting Eq. (F9) into (F6) and multiplying by 2 to include the equal emission and absorption rates, we obtain

$$W_{kk'}^{(\rm ap)}=\frac{\Xi_a^2}{\hbar\beta\lambda_T^2\rho_m s^2}\inf\{1,\exp[-\lambda_T^2(k^2-k'^2)/2]\}\ . \qquad \text{(F10)}$$

Given the expressions such as (F8) and (F10) we can readily construct the collision operator using Eq. (F2). For the purposes of numerical evaluation, it is most convenient to accumulate the values of $C_B(k,k')$ (in the discrete approximation) by performing the assignments

$$\begin{aligned} C_B(k,k')&\leftarrow C_B(k,k')+(\Delta_p/2\pi\hbar)W_{kk'}\ ,\\ C_B(k,k)&\leftarrow C_B(k,k)-(\Delta_p/2\pi\hbar)W_{kk'}\ , \end{aligned} \qquad \text{(F11)}$$

for all values of k and k'. One can implement this procedure in a single subprogram to which a function that evaluates $W_{kk'}$ is passed as an argument, and then invoke this subprogram for each of the processes of interest. A convenient test of the resulting C_B is provided by the principle of detailed balance. It is $C_B f_{\rm eq}=0$, where $f_{\rm eq}$ is an equilibrium (Maxwellian) distribution. The collision operators obtained from Eqs. (F8) and (F10) pass this test.

The effects of the Boltzmann collision operators for these phonon scattering processes on the steady-state characteristics of the RTD are illustrated in Fig. 25. In this calculation the matrix elements for GaAs using the parameters of Fawcett, Boardmann, and Swain (1970) were assumed to hold throughout the structure. The acoustic-phonon scattering has a very small effect on the $J(V)$ curve. The longitudinal-optic phonon scattering processes significantly decrease the peak current and increase the valley current. The initial report of this calculation (Frensley, 1988b) employed a scattering operator for the longitudinal-optic phonons which was one-half of the correct value, due to an algebraic error. Similar calculations have been done by Mains and Haddad (1988b). Kluksdahl *et al.* (1989) and Jensen and Buot (1990) have used a relaxation term to model the inelastic processes.

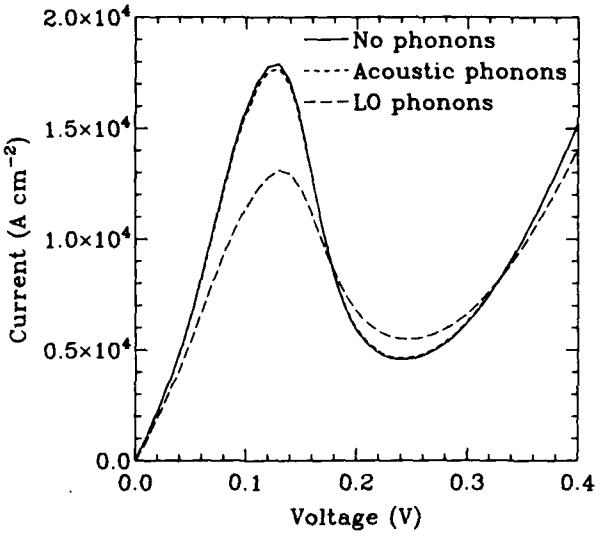


FIG. 25. Effect of phonon scattering processes on the $J(V)$ characteristic of the resonant-tunneling diode, using the Boltzmann collision operator. Scattering by longitudinal-optic phonons significantly reduces the peak current and increases the valley current. The effect of acoustic phonons is nearly negligible. The temperature was 300 K.

APPENDIX G: DEVELOPMENT OF THE DISCRETE WIGNER DISTRIBUTION FUNCTION FOR SIGNAL ANALYSIS

The Wigner distribution function has been found to be useful in the field of signal analysis, where it provides a way to define a time-dependent frequency spectrum (Claasen and Mecklenbräuker, 1980). The notion that a frequency distribution can vary with time is quite intuitive: Consider our usual concept of music as a temporal sequence of notes. But it encounters precisely the same problem with respect to the Fourier uncertainty principle that the notion of a position-dependent momentum distribution does with respect to the quantum-mechanical uncertainty principle. Thus the Wigner distribution may be employed for the same purpose as in quantum mechanics: as a rigorous description that has a simple interpretation in the "classical" regime (in this case, for signals whose frequency spectrum changes slowly).

The relevance of this body of work to the present discussion is that digital signal analysis employs discretely sampled signals that are fully analogous to the discrete models discussed in Sec. VI.A. Many of the mathematical properties (and difficulties) of the discrete Wigner distribution discussed there have already been explored in the context of signal analysis. The purpose of this Appendix is to delineate the parallels between the signal-analysis work and the work reviewed in the body of the present paper.

In the signal-analysis problem, one has a function $\bar{x}(t)$ that has been sampled with an interval T so that only the

values $x(n)=\bar{x}(nT)$ are known for integral n. The sampled signal corresponds to a Schrödinger wave function defined on a spatially discrete basis. The autocorrelation sequence, $\phi_{xx}(m,n)=x_m^* x_n$ (or a statistical average of this quantity, Oppenheim and Schafer, 1975, Chap. 8), corresponds to the density matrix. The Wigner distribution function $f(n,\theta)$, where n represents the time (and corresponds to j) and θ represents the frequency (and corresponds to p), is obtained from the autocorrelation sequence by a transformation similar to Eq. (4.13).

The initial work on the discrete Wigner distribution by Claasen and Mecklenbräuker (1980) used precisely the definition (4.13) (but regarded θ as a continuous variable). They observed that only one-half of the autocorrelation information is employed in this definition, as illustrated in Fig. 21, and noted that, as a consequence, θ is periodic with a period of π rather than 2π. (The corresponding expression in the present work is $N_p\Delta_p=\pi\hbar/\Delta_q$.) In a later work, Claasen and Mecklenbräuker (1983) investigated the consequences of modifying the definition of the discrete Wigner distribution by modifying the kernel of the transformation (4.13) to be something more elaborate than just an exponential function. In particular, they weighted the exponential by a factor very similar to that which appears in Eq. (7.10), used by Mains and Haddad (1988a, 1988c) to weight the potential kernel. Poletti (1988) has further developed this analysis.

If the details of the physical system that produced the signal $x(n)$ are unknown, as is usually the case in signal analysis, the analog of the Liouville equation is also unknown. Thus, in this context, it is natural to try to resolve the problems of the discrete Wigner function by modifying the expression by which it is defined. This approach is complementary (and quite possibly equivalent) to that explored in Sec. VII for modifying the Liouville equation.

REFERENCES

Ando, T., A. B. Fowler, and F. Stern, 1982, Rev. Mod. Phys. **54**, 437.

Apostol, T. M., 1969, *Calculus, Vol. II, Multi-Variable Calculus and Linear Algebra* (Blaisdell, Waltham, MA), Chap. 1.

Aubert, J. P., J. C. Vaissiere, and J. P. Nougier, 1984, J. Appl. Phys. **56**, 1128.

Baranger, H. U., and J. W. Wilkins, 1987, Phys. Rev. B **36**, 1487.

Bardeen, J., 1949, Bell Syst. Tech. J. **28**, 428.

Barker, J. R., D. W. Lowe, and S. Murray, 1984, in *The Physics of Submicron Structures*, edited by H. L. Grubin, K. Hess, and D. K. Ferry (Plenum, New York), p. 277.

Bedeaux, D., K. Lakatos-Lindenberg, and K. E. Shuler, 1971, J. Math. Phys. **12**, 2116.

Berry, M. V., 1977, Philos. Trans. R. Soc. London **287**, 237.

Bjorken, J. D., and S. D. Drell, 1964, *Relativistic Quantum Mechanics* (McGraw-Hill, New York), Sec. 6.3.

Blakemore, J. S., 1982, Solid-State Electron. **25**, 1067.

Broekaert, T. P. E., W. Lee, and C. G. Fonstad, 1988, Appl. Phys. Lett. **53**, 1545.

Büttiker, M., Y. Imry, R. Landauer, and S. Pinhas, 1985, Phys. Rev. B **31**, 6207.

Cahay, M., M. McLennan, S. Datta, and M. S. Lundstrom, 1987, Appl. Phys. Lett. **50**, 612.

Caldeira, A. O., and A. J. Leggett, 1983, Physica A **121**, 587.

Carruthers, P., and F. Zachariasen, 1983, Rev. Mod. Phys. **55**, 245.

Castagné, R., 1985, Physica B **134**, 55.

Champlin, K. S., D. B. Armstrong, and P. D. Gunderson, 1964, Proc. IEEE **52**, 677.

Chang, L. L., L. Esaki, and R. Tsu, 1974, Appl. Phys. Lett. **24**, 593.

Chester, G. V., 1963, Rep. Prog. Phys. **26**, 411.

Claasen, T. A. C. M., and W. F. G. Mecklenbräuker, 1980 (in three parts) Philips J. Res. **35**, 217; **35**, 276; **35**, 372.

Claasen, T. A. C. M., and W. F. G. Mecklenbräuker, 1983, IEEE Trans. Acoust. Speech Signal Process., **ASSP-31**, 1067.

Constant, E., 1985, in *Hot-Electron Transport in Semiconductors*, edited by L. Reggiani, Topics in Applied Physics Vol. 58 (Springer, Berlin), p. 227.

Conwell, E. M., 1967, *High Field Transport in Semiconductors*, Solid State Phys., Suppl. 9 (Academic, New York).

Dahl, J. P., 1981, "Dynamical Equations for the Wigner Functions," Technical University of Denmark preprint.

Davies, E. B., 1976, *Quantum Theory of Open Systems* (Academic, London).

Dickinson, H. W., 1938, *A Short History of the Steam Engine* (Cambridge University, Cambridge), Chap. 6.

Dingle, R., W. Wiegmann, and C. H. Henry, 1974, Phys. Rev. Lett. **33**, 827.

Dresden, M., 1961, Rev. Mod. Phys. **33**, 265.

Duderstadt, J. J., and W. R. Martin, 1979, *Transport Theory* (Wiley, New York), Sec. 8.1.2.

Duke, C. B., 1969, *Tunneling in Solids*, Solid State Phys., Suppl. 10 (Academic, New York).

Eastman, A. V., 1949, *Fundamentals of Vacuum Tubes* (McGraw-Hill, New York).

Fawcett, W., A. D. Boardman, and S. Swain, 1970, J. Phys. Chem. Solids **31**, 1963.

Ferry, D. K., 1980, in *Physics of Nonlinear Transport in Semiconductors*, edited by D. K. Ferry, J. R. Barker, and C. Jacoboni (Plenum, New York), p. 117.

Fetter, A. L., and J. D. Walecka, 1971, *Quantum Theory of Many-Particle Systems* (McGraw-Hill, New York), pp. 59–61.

Feynman, R. P., 1972, *Statistical Mechanics, A Set of Lectures* (Benjamin, Reading, MA), Chap. 2.

Frensley, W. R., 1985, J. Vac. Sci. Technol. B **3**, 1261.

Frensley, W. R., 1986, Phys. Rev. Lett. **57**, 2853; **60**, 1589(E).

Frensley, W. R., 1987a, Phys. Rev. B **36**, 1570; **37**, 10 379(E).

Frensley, W. R., 1987b, Appl. Phys. Lett. **51**, 448.

Frensley, W. R., 1988a, Superlattices and Microstructures **4**, 497.

Frensley, W. R . 1988b, Solid-State Electron. **31**, 739.

Frensley, W. R., 1989a, in *Nanostructure Physics and Fabrication*, edited by M. A. Reed and W. P. Kirk (Academic, Boston), p. 219.

Frensley, W. R., 1989b, Solid-State Electron. **32**, 1235.

Frensley, W. R., and J. H. Luscombe, 1990, unpublished.

Frölich, H., 1967, Physica **37**, 215.

Gordon, J. P., 1967, Phys. Rev. **161**, 367.

Grinberg, A. A., and S. Luryi, 1989, Phys. Rev. B **39**, 7466.

Groenewold, H. J., 1946, Physica **12**, 405.

Haken, H., 1975, Rev. Mod. Phys. **47**, 67.

Heinrich, H., G. Bauer, and F. Kuchar, 1988, Eds., *Physics and*

Technology of Submicron Structures, Vol. 83 of *Solid-State Sciences* (Springer, Berlin).
Heller, E. J., 1976, J. Chem. Phys. **65**, 1289.
Hillery, M., R. F. O'Connell, M. O. Scully, and E. P. Wigner, 1984, Phys. Rep. **106**, 121.
Hockney, R. W., and J. W. Eastwood, 1981, *Computer Simulation Using Particles* (McGraw-Hill, New York), Chap. 10.
Horowitz, P., and W. Hill, 1980, *The Art of Electronics* (Cambridge University, Cambridge).
Hu, G. Y., and R. F. O'Connell, 1987, Phys. Rev. B **36**, 5798.
Iafrate, G. J., H. L. Grubin, and D. K. Ferry, 1981, J. Phys. (Paris) Colloq. C7, Suppl. 10, **42**, 307.
Jacoboni, C., and L. Reggiani, 1983, Rev. Mod. Phys. **55**, 645.
Jauho, A.-P., 1989, Solid-State Electron. **32**, 1265.
Jensen, K. L., and F. A. Buot, 1989a, J. Appl. Phys. **65**, 5248.
Jensen, K. L., and F. A. Buot, 1989b, Appl. Phys. Lett. **55**, 669.
Jensen, K. L., and F. A. Buot, 1990, J. Appl. Phys. **67**, 7602.
Kadanoff, L. P., and G. Baym, 1962, *Quantum Statistical Mechanics* (Benjamin/Cummings, Reading, MA).
Keldysh, L. V., 1964, Zh. Eksp. Teor. Fiz. **47**, 1515 [Sov. Phys.—JETP **20**, 1018 (1965)].
Kluksdahl, N. C., A. M. Kriman, D. K. Ferry, and C. Ringhofer, 1988, IEEE Electron Device Lett. **9**, 457.
Kluksdahl, N. C., A. M. Kriman, D. K. Ferry, and C. Ringhofer, 1989, Phys. Rev. B **39**, 7720.
Kohn, W., and J. M. Luttinger, 1957, Phys. Rev. **108**, 590.
Kreuzer, H. J., 1981, *Nonequilibrium Thermodynamics and Its Statistical Foundations* (Oxford Universty, Oxford/New York).
Krieger, J. B., and G. J. Iafrate, 1986, Phys. Rev. B **33**, 5494.
Kubo, R., 1957, J. Phys. Soc. Jpn. **12**, 570.
Kubo, R., M. Toda, and N. Hashitsume, 1985, *Statistical Physics II, Nonequilibrium Statistical Physics* (Springer, Berlin, 1985), Chap. 2.
Lanczos, C., 1961, *Linear Differential Operators* (Van Nostrand, London), Sec. 3.6, Sec. 3.8, and Sec. 4.16.
Landau, L. D., and E. M. Lifshitz, 1959, *Fluid Mechanics*, translated by J. B. Sykes and W. H. Reid (Pergamon, London), Chap. 1.
Landauer, R., 1957, IBM J. Res. Dev. **1**, 233.
Landauer, R., 1970, Philos. Mag. **21**, 863.
Langmuir, I., and K. T. Compton, 1931, Rev. Mod. Phys. **2**, 123.
Langreth, D. C., 1976, in *Linear and Nonlinear Electron Transport in Solids*, edited by J. T. Devreese and V. E. van Doren (Plenum, New York), p. 3.
Lapidus, L., and G. F. Pinder, 1982, *Numerical Solution of Partial Differential Equations in Science and Engineering* (Wiley, New York), Chap. 2.
Lebowitz, J. L., 1959, Phys. Rev. **114**, 1192.
Lebwohl, P. A., and P. J. Price, 1971, Appl. Phys. Lett. **19**, 530.
Lei, X. L., and C. S. Ting, 1985, Phys. Rev. B **32**, 1112.
Lent, C. S., and D. J. Kirkner, 1990, J. Appl. Phys. **67**, 6353.
Levinson, I. B., 1969, Zh. Eksp. Teor. Fiz. **57**, 660 [Sov. Phys. JETP **30**, 362 (1970)].
Louisell, W. H., 1973, *Quantum Statistical Properties of Radiation* (Wiley, New York).
Luryi, S., 1985, Appl. Phys. Lett. **47**, 490.
Luttinger, J. M., and W. Kohn, 1955, Phys. Rev. **97**, 869.
Mahan, G. D., 1987, Phys. Rep. **145**, 251.
Mains, R. K., and G. I. Haddad, 1988a, J. Appl. Phys. **64**, 3564.
Mains, R. K., and G. I. Haddad, 1988b, J. Appl. Phys. **64**, 5041.
Mains, R. K., and G. I. Haddad, 1988c, "Numerical Considerations in the Wigner Function Modeling of Resonant-Tunneling Diodes," University of Michigan preprint.
Mains, R. K., and G. I. Haddad, 1989, "A New Formulation of the Wigner Function Method for Quantum Transport Modeling," University of Michigan preprint.
Mains, R. K., I. Mehdi, and G. I. Haddad, 1989, Appl. Phys. Lett. **55**, 2631.
Makhoul, J., 1975, Proc. IEEE **63**, 561.
Marland, E. A., 1964, *Early Electrical Communications* (Abelard-Schuman, London).
Mehdi, I., and G. I. Haddad, 1989, in *Nanostructure Physics and Fabrication*, edited by M. A. Reed and W. P. Kirk (Academic, Boston), p. 207.
Messiah, A., 1962, *Quantum Mechanics* (Wiley, New York), Vol. 1, p. 120.
Miller, D. R., and D. P. Neikirk, 1990, unpublished.
Milnes, A. G., and D. L. Feucht, 1972, *Heterojunctions and Metal-Semiconductor Junctions* (Academic, New York), Sec. 9.9.
Morrow, R. A., and K. R. Brownstein, 1984, Phys. Rev. B **30**, 678.
Moyal, J. E., 1949, Proc. Cambridge Philos. Soc. **45**, 99.
Narcowich, F. J., and R. F. O'Connell, 1986, Phys. Rev. A **34**, 1.
Oppenheim, A. V., and R. W. Schafer, 1975, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ).
Oppenheim, I., K. E. Shuler, and G. H. Weiss, 1977, *Stochastic Processes in Chemical Physics: The Master Equation* (MIT, Cambridge, MA) and reprints included therein.
Peierls, R., 1974, in *Transport Phenomena*, Vol. 31 of *Lecture Notes in Physics*, edited by G. Kirczenow and J. Marro (Springer, Berlin), p. 1.
Poletti, M. A., 1988, J. Acoust. Soc. Am. **84**, 238.
Pötz, W., 1989, J. Appl. Phys. **66**, 2458.
Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1986, *Numerical Recipes, The Art of Scientific Computing* (Cambridge University, Cambridge), Sec. 17.1.
Prigogine, I., 1980, *From Being to Becoming. Time and Complexity in the Physical Sciences* (Freeman, San Francisco).
Putterman, S. J., 1974, *Superfluid Hydrodynamics* (North-Holland, Amsterdam), Sec. 50.
Ramo, S., 1939, Proc. IRE **27**, 584.
Ravaioli, U., M. A. Osman, W. Pötz, N. Kluksdahl, and D. K. Ferry, 1985, Physica B **134**, 36.
Reed, M. A., W. R. Frensley, W. M. Duncan, R. J. Matyi, A. C. Seabaugh, and H.-L. Tsai, 1989, Appl. Phys. Lett. **54**, 1256.
Reed, M. A., and W. P. Kirk, 1989, Eds., *Nanostructure Physics and Fabrication* (Academic, Boston).
Reggiani, L., 1985, in *Hot-Electron Transport in Semiconductors*, Vol. 58 of *Topics in Applied Physics*, edited by L. Reggiani (Springer, Berlin), p. 7.
Reichl, L. E., 1980, *A Modern Course in Statistical Physics* (University of Texas, Austin, TX).
Reynolds, T. S., 1983, *Stronger than a Hundred Men, A History of the Vertical Water Wheel* (Johns Hopkins University, Baltimore), Chap. 4.
Ringhofer, C., D. K. Ferry, and N. C. Kluksdahl, 1989, Transport Theor. Stat. Phys. **18**, 331.
Roache, P. J., 1976, *Computational Fluid Dynamics* (Hermosa, Albuquerque, NM).
Scully, M. O., and W. E. Lamb, Jr., 1967, Phys. Rev. **159**, 208.
Selberherr, S., 1984, *Analysis and Simulation of Semiconductor Devices* (Springer, Vienna).
Shockley, W., 1938, J. Appl. Phys. **9**, 635.
Shockley, W., 1949, Bell Syst. Tech. J. **28**, 435.
Slater, J. C., 1949, Phys. Rev. **76**, 1592.

Sollner, T. C. L. G., W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, 1983, Appl. Phys. Lett. **43**, 588.
Stone, A. D., and A. Szafer, 1988, IBM J. Res. Dev. **32**, 384.
Szafer, A., and A. D. Stone, 1989, Phys. Rev. Lett. **62**, 300.
Tolman, R. C., 1938, *The Principles of Statistical Mechanics* (Oxford University, Oxford; republished by Dover, New York, 1979).
Tsu, R., and L. Esaki, 1973, Appl. Phys. Lett. **22**, 562.
Visscher, P. B., 1988, *Fields and Electrodynamics, A Computer-Compatible Introduction* (Wiley, New York).
Visscher, P. B., 1989, Comput. Phys. **3**(2), 42.
Webb, R. A., 1989, in *Nanostructure Physics and Fabrication*, edited by M. A. Reed and W. P. Kirk (Academic, Boston), p. 43.
Wigner, E., 1932, Phys. Rev. **40**, 749.
Wigner, E. P., 1971, in *Perspectives in Quantum Theory*, edited by W. Yourgrau and A. van der Merwe (MIT, Cambridge, MA), p. 25.
Wilkinson, J. H., 1965, *The Algebraic Eigenvalue Problem* (Oxford University, London), Chap. 2, Sec. 13.
Wingreen, N. S., and J. W. Wilkins, 1987, Bull. Am. Phys. Soc. Ser. II **32**, 833.
Wolf, E. L., 1985, *Principles of Electron Tunneling Spectroscopy* (Oxford University, New York).
Woolf, H. B., 1981, Editor in Chief, *Webster's New Collegiate Dictionary* (Merriam, Springfield, MA), p. 200.
Yennie, D. R., 1987, Rev. Mod. Phys. **59**, 781.
Zwanzig, R., 1964, Physica **30**, 1109.

APPENDIX II
NANO2D: A TWO-DIMENSIONAL HETEROSTRUCTURE DEVICE MODELING PROGRAM

NANO2D : A TWO-DIMENSIONAL HETEROSTRUCTURE DEVICE MODELING PROGRAM

Ann M. Bouchard and James H. Luscombe
Central Research Laboratories
Texas Instruments Incorporated
Dallas, Texas 75265

I. Summary

NANO2D is a general two-dimensional device simulation code which obtains the self-consistent potential energy surface defined by the conduction band minimum for a wide class of two-dimensional III-V semiconductor heterostructure devices. It does so by solving a two-dimensional nonlinear Poisson equation, utilizing a zero-current, local thermodynamic equilibrium approximation for the carrier density. By "two-dimensional" we mean that the user can specify, in addition to an arbitrary sequence of epitaxial growth layers in the vertical direction, an arbitrary *lateral* variation of material composition and doping, such as might be achieved by epitaxial regrowth techniques. In addition, the user can specify the lateral bias across Ohmic emitter and collector contacts as well as voltages applied to a back Ohmic contact and one or more top Schottky gates.

NANO2D is implemented on the CRL VAX system. It may be used to investigate and optimize heterolayer design in high electron mobility transistors (HEMTs) and in lateral resonant tunneling devices (LRTDs) which utilize split-gate top contacts. It may also be used to explore epitaxial regrowth schemes which enhance electron confinement, and to model the effects of the application of different gate and back contact voltages on the electron potential energy surface.

The current implementation explicitly accomodates AlGaAs and InGaAs compounds. However, the code's capability will be generalized to *all* III-V compounds in the forthcoming upgrade.

II. Device Structure

In order to demonstrate the capabilities of the NANO2D code, and to provide users with some sense of what sort of input results in what sort of output, the bulk of this document will be given in the context of an example device, shown schematically in Fig. 1. This particular structure could be a test design for a LRTD. The idea in this design is to confine electrons vertically in a two-dimensional electron gas (2DEG) in the InGaAs layer and create an electrostatic double barrier potential in the lateral direction for lateral resonant tunneling. The two Schottky gates on the top of the structure will control the height of the lateral double barriers. Regrowing n+ doped material outside of the gates is expected to enhance the height to width ratio of the double barrier potential in the InGaAs layer.

Fig. 1(a) shows a top view of the device and indicates the position of emitter and collector contacts, gates, and regrowth regions. Panel (b) shows a vertical cross section along the dashed line in panel (a). This device consists of six layers of semiconducting material, although there is no practical limitation on the number of layers which can be modeled. The top layer contains five lateral regions of GaAs with different doping concentrations. The second layer also contains five

regions, but with variations in the material composition as well as the doping. The thickness of each layer, and the width, composition, and doping in each region within each layer is specified by the user. The position of the left and right edges of up to five Schottky gates is also user-specified.

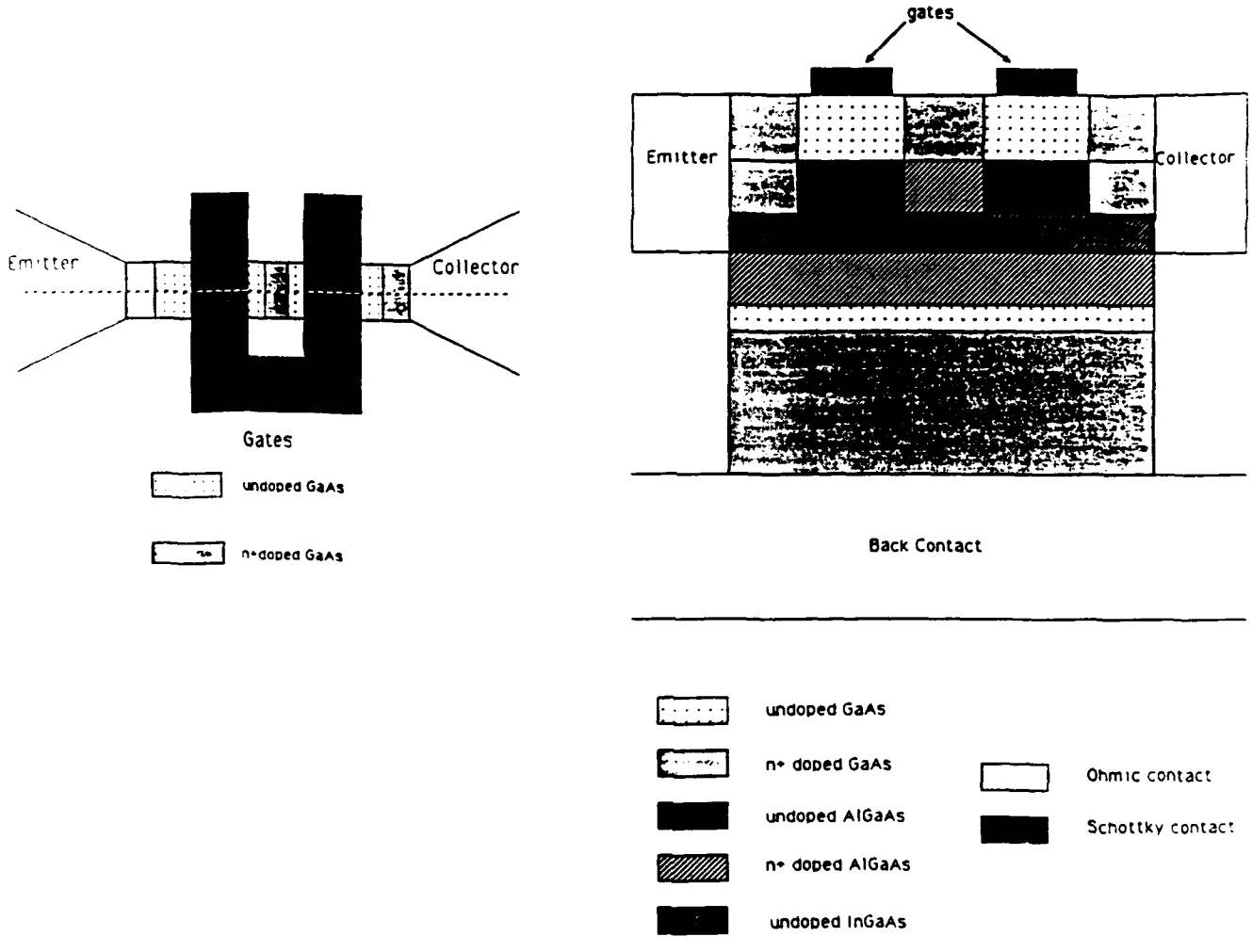


Fig. 1 Generic device structure for NANO2D

III. Using NANO2D

A. The NANO2D Command File

NANO2D must be run within the VAX environment. Since it takes on the order of five CPU minutes to run, users are advised to submit the job to a batch queue using the SUBMIT command. The command file which compiles, links, and runs NANO2D, and which supplies the input, is found in SCL:[BOUCHARD.RELEASE]NANO2D.COM. Users should copy this file into their own directories and make changes to input in their own versions. Throughout this technical report references are made to specific lines of the NANO2D.COM file. To aid the reader, the Appendix contains a copy of this file, with input corresponding to the example device of Fig. 1.

The first change which should be made to [your_directory]NANO2D.COM is the default directory. The second line of the file should read

$ SET DEFAULT [your_directory]

The string "your_directory" should be replaced by the name of the directory where output files are to be written. Aside from the NANO2D.LOG file, which contains a summary of the device parameters and other useful information, the only output files are HP plotter files which are printed automatically by the NANO2D.COM file and may be deleted as soon as printing is completed.

B. Input

The eighth line of NANO2D.COM issues the command to run the NANO2D program. The lines that follow immediately after the RUN command are input statements. The following discusses the required input. **Input parameters appearing on the same line should be separated by blank space, not by commas, periods, or any other punctuation marks.**

(1) *title*
The 80-character line immediately following the RUN command is reserved for user comments. It may be used to identify or "title" the particular set of input parameters, and will appear as a title on graphical output. It may contain any combination of alphanumeric characters, punctuation, and white space.

(2) Device Structure Input

(a) *vmesh lmesh temp*
The line immediately following the title must contain three real values, the vertical mesh spacing (in nm), *vmesh*, the lateral mesh spacing (in nm), *lmesh*, and the temperature (in K), *temp*. Any comments to the right of these three real values are ignored by the program. **We recommend that *vmesh* be chosen such that no semiconductor layer contains fewer than five mesh points, and that *lmesh* be chosen such that no lateral region contains fewer than eight mesh points. This will ensure a potential energy surface which varies smoothly in both dimensions.**

(b) The next section of input contains the structure information, layer by layer, and region by region within a layer. The specification of a single layer requires one line of input for each region of material in the layer. (e.g. The top layer of Fig. 1 requires five lines of input, whereas the third layer requires only one line, as shown in the Appendix.) Specifically, the input for a layer consists of (i) one line for the first (left-most) region; (ii) if necessary, lines for the remaining regions, in order from left to right; and (iii) NEXT, to signal to continue to the next layer. The form of the input (i), (ii), and (iii) are given in the following:

(i) *lthick width dope* material composition
The first line of input for one semiconductor layer contains the layer thickness (in nm), *lthick*, the width (in nm) of the first (left-most) region in that layer, *width*, the doping concentration (in cm^{-3}) of the first region, *dope*, and the material composition in the first region. (The format of the material composition is discussed in sub-section (d).)

(ii) *width dope* material composition

The second and subsequent lines for the same layer just contain the *width*, *dope*, and material composition of the respective region within that layer. **Note that *lthick* is not included in these input lines, since the thickness for all the regions in the layer is specified in (i).**

(iii) NEXT

When the last (right-most) region of the layer has been specified in this way, it is followed by a line containing the word **NEXT**, meaning go on to the next layer. It must be in all capital letters with no punctuation.

(c) END

When all layers have been entered **(the last layer must also end with NEXT)**, the following line contains the word **END**, indicating the end of semiconductor layer input. It also must be in all capital letters with no punctuation.

(d) The input format for the material composition is easiest described through a series of examples:

(i)	GaAs	is indicated by	Ga 1.00
(ii)	InAs	is indicated by	In 1.00
(iii)	$Ga_{0.7}Al_{0.3}As$	is indicated by	Ga 0.70 Al 0.30
(iv)	$In_{0.15}Ga_{0.85}As$	is indicated by	In 0.15 Ga 0.85
(v)	$In_{0.05}Al_{0.2}Ga_{0.75}As$	is indicated by	In 0.05 Al 0.20 Ga 0.75

The pattern here is straightforward. The Group-III element is followed by its mole-fraction. As long as there is additional input to the right, the program continues to read it. Currently, it is assumed that the Group-V element is As. In the forthcoming upgrade, however, other Group-V elements will also be allowed. In the newer version, if no Group-V element appears in the input line, then As will be assumed. If, however, P or Sb, is desired, that is indicated with the appropriate chemical symbol to the right of the Group-III input, as in the following examples:

(vi)	InP	is indicated by	In 1.00 P
(vii)	AlSb	is indicated by	Al 1.00 Sb
(viii)	$In_{0.25}Ga_{0.75}P$	is indicated by	In 0.25 Ga 0.75 P

Note: It is important not to have extra characters or comments to the right of the structure portion of the input (part (2)), as the program will read it and attempt to interpret it as additional input data.

(3) Boundary condition information

(a) *f_lev_pin*

The line immediately following the 'END' of the structural input must contain one real number, *f_lev_pin*, the energy value (in eV) of the Fermi-level pinning of the top layer of semiconducting material. For GaAs this number is usually taken to be 0.7 eV, half of the band gap energy. If the top layer is something other than GaAs, then a *f_lev_pin* value equal to half the band gap of the surface material is a reasonable choice.

(b) The next section of input contains information about the top contacts, or Schottky gates. The first line simply specifies the number of gates. The lines that follow give details about each gate, one line of input per gate, ordered from left to right.

(i) *ngates*

The next line contains an integer, *ngates*, the number of gates. The current version of the program supports up to five gates. **Note: If you wish to run with no gates, be sure to read Section V, "What Else Users MUST Know Before Modeling a Device."**

(ii) *lpos rpos voltage Schot_bar*

The following ngate lines each contain four real numbers characterizing each of the gates. *Lpos* specifies the position of the left edge of the gate; *rpos* specifies the position of the right edge of the gate. Both are measured in nm from the left boundary of the device. *Voltage* is the voltage applied to the gate in volts, and *Schot_bar* is the height (in volts) of the Schottky barrier formed at the interface between the gate and the top semiconductor. The gates should be ordered from left to right. e.g. If ngates = 3, then the left-most gate should be entered first, the middle gate should be entered second, and the right-most gate should be entered last.

(c) Ohmic contact voltages

(iii) *v_emit v_collect*

The next line contains two real numbers, *v_emit* and *v_collect* (in volts), the voltage applied to the emitter (left) contact and collector (right) contact, respectively.

(iv) *v_back*

The next line contains a real number *v_back* (in volts), the voltage applied to the back contact.

Remarks:

(1) The contact voltages should be specified with respect to some ground. For example, one could specify *v_emit* =0.0 (ground) and specify *v_collect* and *v_back* with respect to *v_emit*.

(2) The Schottky barriers should be positive valued and indicate the height to which the Fermi level of the semiconductor is pushed up with respect to the conduction band of the metal gate.

(3) If a bias is to be applied from the emitter to the back contact, then at least one larger-band-gap material layer must isolate the back contact from the emitter and collector regions. See Section V, "What Else Users MUST Know Before Modeling a Device."

(4) Output Options

(a) *flag*

The next line contains an integer *flag*. If *flag* ≤ 0, then two figures are generated: a plot of a lateral slice of the potential energy surface, and a plot of a vertical slice down the center of the potential energy surface. If *flag* > 0, then three figures are generated: the lateral slice, the center slice, and a plot of the full potential energy surface.

(b) *depth*

The next line contains the *depth* (in nm) below the surface of the structure where the lateral slice plot is to be generated.

(c) *pen_speed*

The last line of input contains an integer, the HP plotter pen speed. For most cases "2" is a good choice. For publications or foils, we recommend a slower speed, "1" or even "0".

Remark:

If *flag* ≤0, then it is a good idea to put an exclamation point (!) in front of the last statement of the NANO2D.COM file. It would then read:

!$ pmhp hp7550a.dat;-2

This prevents printing a version of the HP-plotter file left over from a previous run.

C. Modeling a Device

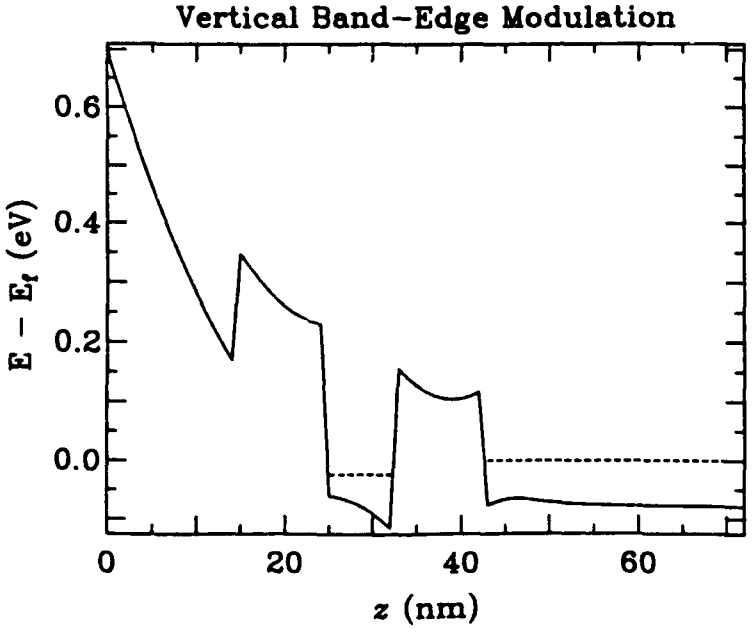
When the input has been changed to specify the structure of the device of interest, simply SUBMIT the NANO2D.COM file to a batch queue. Within ten or fifteen minutes wall-clock time the graphical output will be printed on the HP plotter in the VAX printer room in the Research West building. If output is desired in the Research East building, the last three commands in the NANO2D.COM file must be changed from "pmhp" to "prhp". For foils in Research West, the command is "pfmhp", and for foils in Research East, it is "pfrhp". The lateral slice plot, the center slice plot, and if flag > 0 the full potential surface are printed on the designated printer when the program has finished running.

IV. Output

The NANO2D.LOG file contains a report of the various input parameters of the run and other diagnostic statements. Most of the output is in graphical form.

A. What the Output Looks Like

The first figure output is a constant-z slice of the potential as a function of the lateral position x. The z-position of the slice is specified by the variable *depth*, as discussed in the last section. Fig. 2 illustrates this output for the example device of Fig. 1, with a *depth* of 27.0 nm which is just at the top of the InGaAs layer. The dashed lines (these will be red in the actual output from the program) indicate the position of the Fermi level in equilibrium with the emitter (to the left) and the collector (to the right).



The second figure plotted is a constant-x slice of the potential down the center of the device. Fig. 3 shows the center plot for the example device.

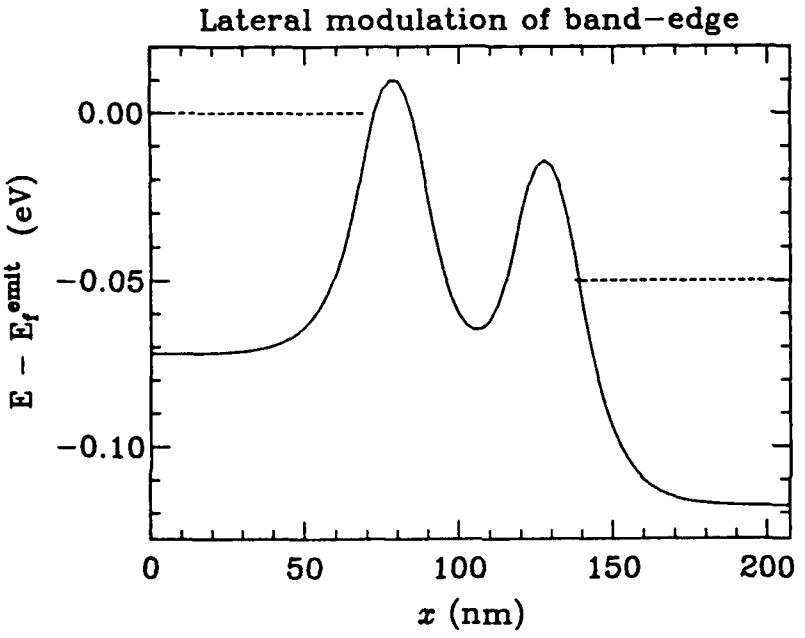


Fig. 3

The lateral position for the slice is half-way between the emitter and collector, i.e. half of the x dimension of the device. For a device with two gates symetrically placed, this gives a slice mid-way between the two gates. For some other number of gates, one may prefer other vertical slice positions. See Section VII, "User Feedback" to report suggested changes.

The dashed lines (again, these will be red in the actual output) show the position of the Fermi level. Since it is assumed that the back contact is isolated from the emitter/collector region of the device by an AlGaAs barrier near the bottom of the device (e.g. layer 4 of the example device serves this purpose), then by default the Fermi level is drawn to either side of this layer. If no such layer exists in your model, and some alternative Fermi-level plotting is needed, again see section VII, "User Feedback" to request changes.

If the variable *flag* is greater than zero, then a third figure is also plotted, the full two-dimensional potential energy surface as a function of both x and z. This plot is shown, for the example device, in Fig. 4. The z-axis is reversed in this figure to provide the best view

of the surface. Users must keep in mind that although in this figure the origin is in the *lower* left corner of the device, in the input, the origin is in the *upper* left corner.The contours (red in the actual output) indicate the position of the Fermi level in equilibrium with the emitter (near edge), the collector (far edge), and the back contact (to the right).

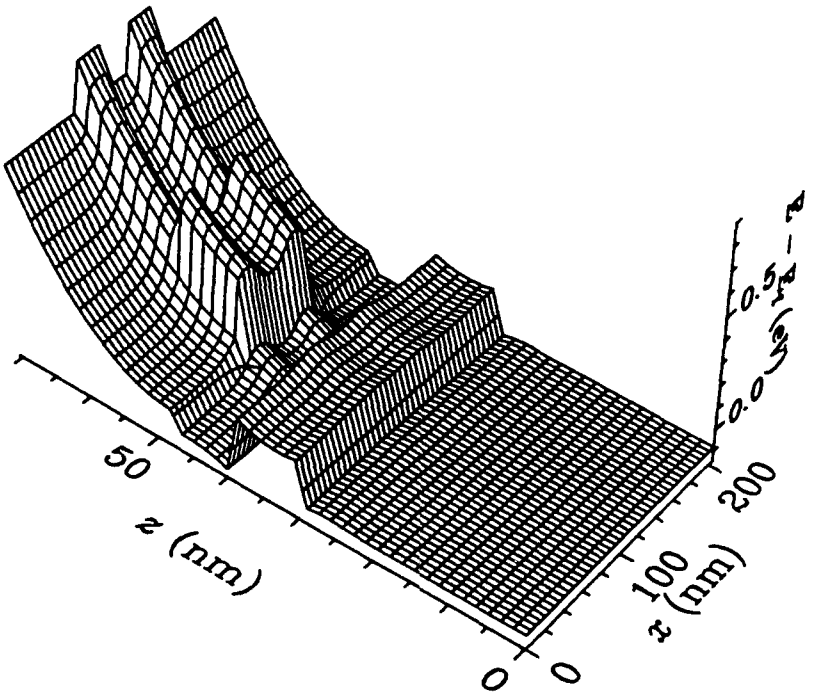


Fig. 4 Complete band edge surface for structure listed in the Appendix

B. Understanding the Output

In order to gain some sense of what these figures mean, look first at the 2-D conduction band profile (Fig. 4) and compare it with the structure of the device in Fig. 1. The top of the device is to the upper left, where the potential bends up due to Fermi-level pinning. The two barriers which originate at the top of the structure are due to a voltage applied to the Schottky gates. The barrier from z=30 to 40 nm is the AlGaAs layer (layer 4), and the two tall bumps at about z=50 nm are due to the undoped AlGaAs regions of layer 2.

The first thing to notice is that, in general, regions of AlGaAs have relatively high potential energy, InGaAs has a very low potential energy, and the GaAs regions are somewhere in between. This is because of the difference in band gap of the three types of materials. Secondly, note the big dip in potential energy in the region between the undoped AlGaAs regions of layer 2. Even though all three regions in the middle of layer 2 are AlGaAs, the heavily doped region has a lower potential energy than the undoped regions. This device design exploits such effects by including regions of regrown heavily doped GaAs on the outside of the gates. This enhances the deepening of the InGaAs layer on the outside, to increase the height to width ratio of the double barriers.

Now that we have a global idea of what is going on in the device, from looking at the 2-D potential surface, we can focus on what is going on in some particular area of interest. In the example device, that region of interest is the InGaAs layer, where we hope to have a good

double-barrier potential set up with the Fermi level above the conduction band both to the left and to the right of the barriers. The slice-plot (Fig. 2) shows clearly that this has been achieved, and allows a clearer view of the lateral shape of the potential. The center-plot (Fig. 3) provides a "side view", to allow us to determine how isolated the InGaAs layer is from the back contact, and from the surface of the device.

V. What Else Users MUST Know Before Modeling a Device: "Boot-Strap" Approach to Modeling a Device

NANO2D solves for the self-consistent conduction-band profile using the "zero-current" approximation. This approximation is valid for a device under bias only if the contacts are sufficiently isolated from each other by potential barriers. We recommend a "boot-strap" approach to modeling a device under bias: First, model the device with zero volts at all three contacts. Then add gates, and adjust the gate voltages and back contact voltage until you are satisfied that the emitter and collector are isolated by electrostatic potential barriers. *Then*, you are ready to apply an emitter-collector bias.

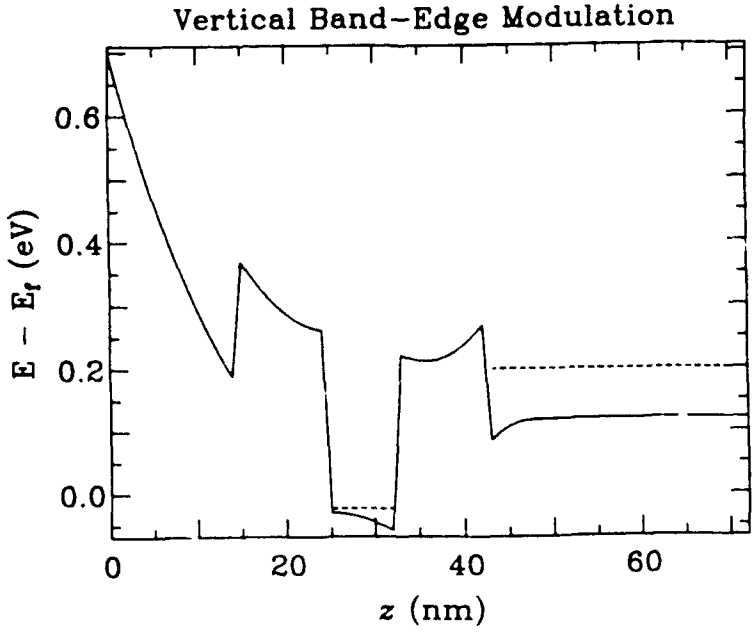
If two contacts are not sufficiently isolated and there is a potential difference between them, the program will work in an unpredictable way. Either it will not converge, or it will converge to a a solution which is likely incorrect. It is *very* important to make sure the contacts are isolated before applying a bias.

It follows, therefore, that this program is not appropriate for modeling a device under bias with *no* gates unless there is a large-band-gap material region separating the emitter from the collector.

VI. Some Helpful Hints

A. Biasing the Back Contact

Applying a voltage to the back contact which is *negative* with respect to the emitter results in the energy band near the back contact being pulled *up*. To illustrate this effect, Fig. 5 shows the center plot for the same example structure as in Fig. 3, but with a *v_back* = -0.2 V.



We have found that this negative bias tends to pull electrostatic barriers set up by the gates deeper into the device. Note in Fig. 6 how much higher the barriers rise above the Fermi level with -0.2 V applied to the back contact, than in the 0.0 V case of Fig. 2.

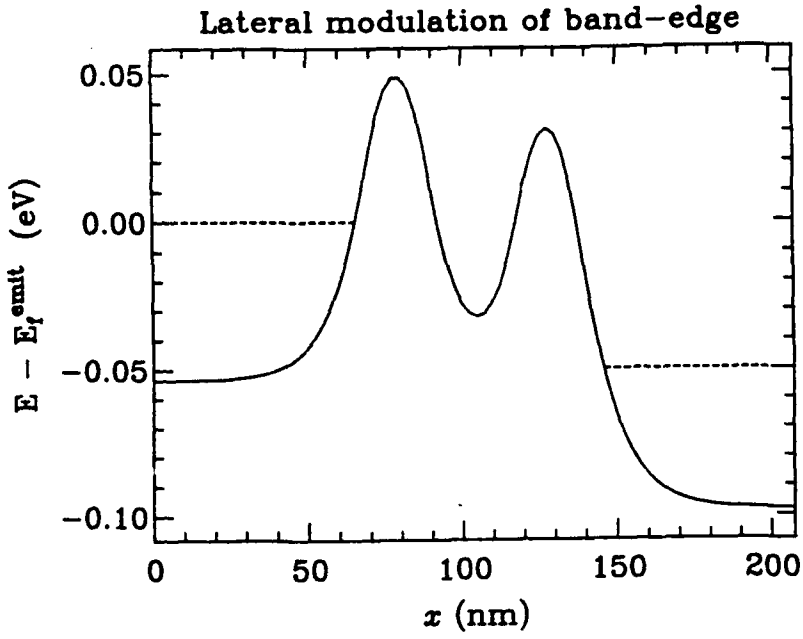


Fig. 6

This may be important for isolating the emitter from the collector with electrostatic barriers. If, for example, the barriers were entirely below the Fermi level without any back-contact voltage, the application of some negative voltage may pull the barriers up sufficiently to isolate the emitter and collector, so that the device could then be modeled with an emitter-collector bias.

B. Surface Depletion

Fermi-level pinning of 0.7 eV results in depletion of the surface to about 25 nm in GaAs. Any lateral pathway through GaAs where current is to flow must be more than 25 nm deep, or it will be depleted. An InGaAs pathway may be placed somewhat less than 25 nm and still not be depleted.

C. One Job at a Time

Each of the graphical output files is named HP7550A.DAT, and a single job creates two or three versions of the same filename. If more than one job is simultaneously creating two or three versions of the same filename all in the same directory, it is difficult to sort out which output

figures belong to which job. We recommend that not more than one NANO2D job be run in the same directory at the same time.

D. Publication-Quality Graphics

The recommended vertical and lateral mesh spacings may be unsuitable for publication-quality graphics. The resolution of the full potential energy surface, after reduction to publication size, would likely be poorer than desired. For making publication-quality figures, a larger mesh spacing may be in order. However, we still strongly recommend a small mesh spacing for every-day use, to ensure accurate simulations. (It goes without saying that one should compare the results of the larger-mesh simulation with those of the smaller-mesh, before publishing.)

VII. User Feedback

Please forward any comments, observations of bugs, or requests for added functionality to James H. Luscombe, 995-6968, MS 154, VAX RESBLD::LUSCOMBE.

VIII. Acknowledgment

The development of NANO2D was supported by ONR Contract No. N00014-89-C-0091, Development of Few-Electron Lateral Resonant Tunneling Semiconductor Devices.

Appendix

```
$ ASSIGN NL: SYS$PRINT
$ SET DEFAULT [bouchard.release]
$ set noverify
$ pvi
$ set verify
$ PASCAL nano2d
$ DI3LOAD nano2d,[FRENSLEY.GRAPHICS]WRFPLOT.OLB/library share
$ purge nano2d.*
$ RUN nano2d
Lateral Resonant Tunneling Device
1.6 5.0 300.0       VERT MESH SPACING, LAT MESH SPACING
(NM),TEMPERATURE(K)
15.0  70.0 2.0e18  GA 1.00
    20.0 1.0e10  GA 1.00
    30.0 2.0E18  GA 1.00
    20.0 1.0E10  GA 1.00
    70.0 2.0E18  GA 1.00
NEXT
10.0  70.0 2.0E18  GA 1.00
    20.0 1.0E10  GA 0.75 AL 0.25
    30.0 2.0E18  GA 0.75 AL 0.25
    20.0 1.0E10  GA 0.75 AL 0.25
    70.0 2.0E18  GA 1.00
NEXT
 8.0 210.0 1.0e10  GA 0.90 IN 0.10
```

```
NEXT
10.0  210.0  2.0E18   GA  0.75  AL  0.25
NEXT
 5.0  210.0  1.0E10   GA  1.00
NEXT
25.0  210.0  2.0E18   GA  1.00
NEXT
END
0.7                  f_lev_pin (eV)
2                    number of gates
 72.0  88.0  -0.2  0.7  left and right edges of gate, voltage, and Schottky_bar
122.0 138.0  -0.2  0.7  left and right edges of gate, voltage, and Schottky_bar
0.0  0.05              emitter and collector voltages
0.0                   back contact voltage
1                    flag > 0, 3D plot; flag <= 0, just center slice
25.0                  location of slice to be plotted
1                    pen speed (for hp plotter)
$ pmhp hp7550a.dat;
$ pmhp hp7550a.dat;-1
$ pmhp hp7550a.dat;-2
$ del nano2d.map;*
$ del nano2d.lis;*
$ EXIT
```

APPENDIX III
TUNNELING SPECTROSCOPIC STUDY OF FINITE SUPERLATTICES

Tunneling spectroscopic study of finite superlattices

R. J. Aggarwal,[a] M. A. Reed,[b] W. R. Frensley,[c] Y.-C. Kao, and J. H. Luscombe
Central Research Laboratories, Texas Instruments Incorporated, Dallas, Texas 75265

(Received 15 February 1990; accepted for publication 20 June 1990)

We present a tunneling density of states study of the transition from a superlattice miniband to a sequential coupled well structure. We have observed by tunneling spectroscopy the eigenstates of a finite superlattice system far below the Stark localization threshold. The transition from an indistinguishable miniband to a coupled well structure is experimentally found to be 2.5 meV $< W$(miniband width)$/n($# periods$) <$ 10.5 meV.

Semiconductor superlattices have received renewed interest for the design and fabrication of novel electronic structures utilizing perpendicular transport. A central issue for the design, utilization, and analysis of superlattice structures is the nature of the electronic states. In weakly coupled superlattices it has been shown[1] that the perpendicular transport proceeds via sequential tunneling, whereas under the proper conditions a miniband forms.[2-4] We present here a tunneling density of states study of the transition of a finite superlattice from a superlattice miniband to a coupled well structure.

A generic superlattice tunnel diode structure[5] was utilized to study the density of states in a series of superlattices. Figure 1 shows a self-consistent band diagram at resonant bias (a), along with the experimental current (I) [and conductance (G)] versus voltage (V) characteristics (b), of the type of structures investigated in this study. This specific example is a structure identical to the initial work of Davies *et al.*[5] The band diagram is determined from a self-consistant finite temperature Thomas–Fermi zero-current calculation,[6] with the superlattice structure determined from an envelope function calculation superimposed. When the top of the first collector miniband crosses the bottom of the available emitter electron supply, a decrease in current occurs due to the requirement to conserve both energy and momentum. This is defined as the resonant (peak) voltage. It should be emphasized that realistic band diagrams are necessary for an accurate understanding of resonant effect.

Table I illustrates the series of superlattice structures investigated. Structure S1 was identical to that of Davies *et al.*[5] The remaining samples consisted of a Cr-doped semi-insulating GaAs substrate, a 0.5 μm undoped GaAs buffer, a 1.0 μm 1×10^{18} cm^{-3} n^+-GaAs bottom contact, a 420 Å 1×10^{17} cm^{-3} (last 20 Å undoped) GaAs contact to superlattice transition region, a superlattice/tunnel barrier/superlattice region symmetric about the tunnel barrier, a 400 Å 2×10^{18} cm^{-3} GaAs top contact, and an InGaAs top nonalloyed ohmic contact. To study the effects of contact doping, S3 had symmetric 400 Å 1×10^{17} cm^{-3} contact regions adjacent to the superlattices. S5 was identical to S4, except that the bottom superlattice was replaced with bulk GaAs (though the doping modulation was identical). Structural parameters were verified by cross-section transmission electron microscopy, and photoluminescence of nominally identical superlattices (grown without doping and contact structures) was used to verify superlattice

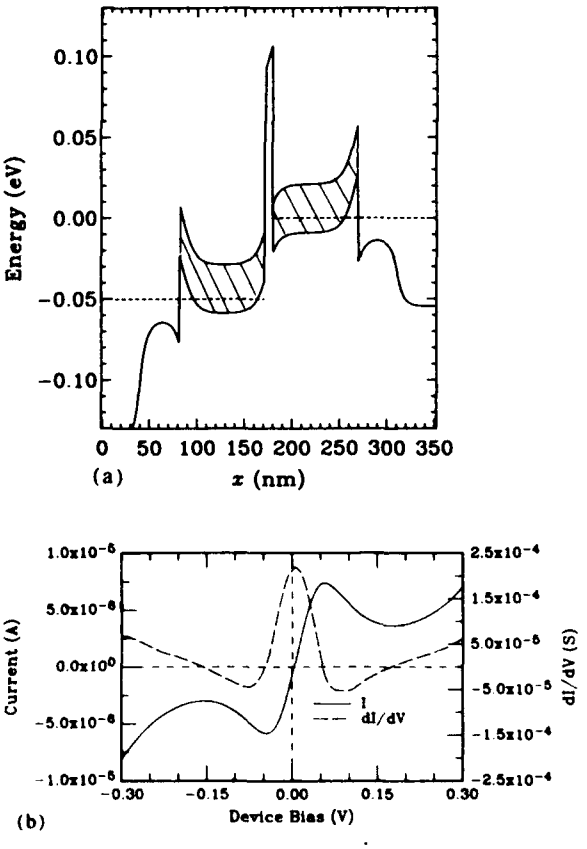


FIG. 1. (a) Self-consistent Γ-point energy band vs epitaxial dimension of sample S1 at resonant bias. The hatched regions denote the 25-meV-wide lowest superlattice minibands and the dotted lines the Fermi level. The structure is identical to that reported by Davies *et al.* (Ref. 15) $T = 4.2$ K. (b) Experimental current (solid) and conductance (dashed) vs voltage characteristics of S1. $T = 4.2$ K.

[a]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.
[b]New address: Department of Electrical Engineering, Yale University, P. O. Box 2157 Yale Station, New Haven, CT 06520-2157.
[c]New address: Erik Jonsson School of Electrical Engineering and Computer Science, University of Texas at Dallas, P. O. Box 830688, Richardson, TX 75083-0688.

 0003-6951/90/330707-03$02.00

TABLE I. Summary of the superlattice tunneling structures investigated. $E_{SL,min} - E_c$ denotes the energy of the bottom of the first miniband (in meV), referenced to GaAs. W denotes the width (in meV) of the first miniband. The superlattice minibands were calculated using an infinite envelope function approximation. $E_{F,SL} - E_c$ denotes the Fermi energy of the superlattice (in meV), referenced to GaAs.

Sample	d_{GaAs}/d_{AlGaAs} (Å)	$E_{SL,min} - E_c$ (meV)	W (meV)	$E_{F,SL} - E_c$ (MeV)
S1	60/30	53	30	62
S2	40/50	90	25	96
S3	49/14	45	105	57
S4	40/10	43	190	55
S5	40/10 (asymmetric)	43	190	55

band gap and aluminum content. Mesas as small as $4(\mu m)^2$ were fabricated using standard contact lithography processing.

The superlattice structure S1 is presented to compare to previous work.[5] Structures S2–4 are also ten period superlattices, designed such that the superlattice miniband widths span the available range in the conduction band. The GaAs wells of these superlattices were doped at 1×10^{17} cm^{-3}, the $Al_{0.23}Ga_{0.77}As$ barriers were nominally undoped, and the tunneling barrier was kept fixed at 100 Å of $Al_{0.23}Ga_{0.77}As$. S2 has the same approximate miniband width as S1, with the second miniband "virtual" only. S3 is designed to have the same approximate superlattice energy centroid as S2, with a factor of 4 larger miniband width. S3 and S4 have the same miniband minimum, with S4 having almost a factor of 2 larger miniband width than S3. In addition, the top of S4's first miniband is "virtual." S5's superlattice is identical to S4, except the asymmetry allows one to investigate injection into a superlattice from a 3D system, and vice versa.

The superlattice Fermi levels were calculated by assuming free electrons in the transverse directions and Bloch states for the vertical direction. The Fermi level was then inferred as the chemical potential which leads to a miniband-occupied carrier density corresponding to the average carrier concentration of the sample.[7] It should be noted that determination of the superlattice Fermi level in general produces a higher Fermi level than that for a bulk system of the same density.

The I-V and G-V characteristics of samples S1 and S2 are very similar, exhibiting well-defined negative differential resistance (NDR) at low temperature [with peak-to-valley (P/V) current ratios as high as 2:1 for S1, 2.4:1 for S2]. NDR is observable (P/V 1.3:1) at room temperature in S2, and an inflection is clearly evident at room temperature in S1.[8] Aside from the major resonance (Fig. 1), there is no apparent additional structure in the conductance greater than the 1 mV (i.e., 12 K) experimental resolution for either S1 or S2, at a sample temperature of 4.2 K (immersed).

We now experimentally increase the superlattice miniband from 25 to 105 meV and examine the vertical transport. Figure 2 shows the low-voltage I-V and G-V characteristics of S3 at 4.2 K. The $\pm \sim 120$ mV major peak

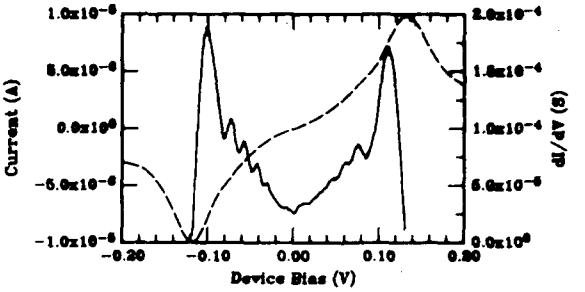


FIG. 2. Low-voltage I-V (solid) and G-V (dashed) characteristics of sample S3 (105-meV-wide superlattice miniband) at 4.2 K. The $\pm \sim 120$ mV major resonances corresponds to the alignment line-up of the first minigap with the emitter.

corresponds to the line-up of the first minigap with the emitter. A series of peaks on the low bias side of the major peak is apparent. Note that these biases correspond to electric fields well below that expected for Stark localization.[9,10] The condition for Stark localization of a superlattice is $eEd > W$, where E is the applied electric field, d is the superlattice period, and W the width of the miniband under consideration. At the biases considered here, the Stark splitting is < 10 meV, compared to a miniband width of 105 meV. The "subresonant series" starts to degrade above 20 K, and is unobservable (except for the highest subresonance peak) above 50 K.

Figure 3 shows the I-V and G-V characteristics of a superlattice miniband experimentally increased to 190 meV (S4), keeping the number of superlattice periods constant. The subresonance series is very pronounced; higher bias peaks are evident even at room temperature. Assuming that the structure is due to the finite extent of the superlattice, we calculate the single electron transmission coefficient of the ten-period superlattice/coupled quantum well system, and map these ten resonant peaks onto the self-consistent band structure. Figure 4 shows the calculated resonant crossings of the collector finite superlattice transmission peaks with the emitter Fermi level, compared with the experimental resonant peaks. The calibration of the top resonance is determined by the number of periods in the finite superlattice, and the low peak cutoff is determined from the superlattice Fermi level. The agreement between calculated and experimental peak position is qualitatively (a $V^{1/2}$ behavior) and quantitatively good. Like-

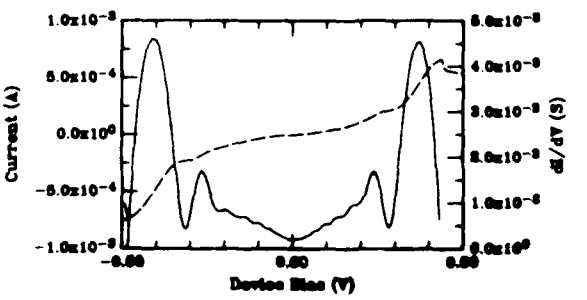


FIG. 3. Low-voltage I-V (dashed) and G-V (solid) characteristics of sample S4 (190-meV-wide superlattice miniband) at 10 K.

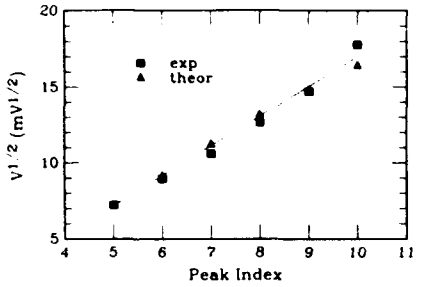


FIG. 4. Experimental (square) and theoretical (circle) resonant crossings of the collector finite superlattice transmission peaks with the emitter superlattice Fermi level. The calculated resonant crossings were determined from mapping the finite superlattice transmission peaks onto the self-consistent band structure and determining the bias at which they cross the emitter Fermi level.

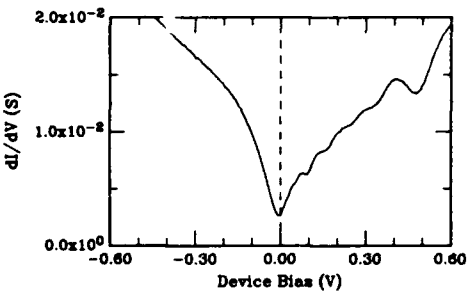


FIG. 5. *G-V* characteristics of sample S5 (S4 with one superlattice replaced with bulk GaAs). $T = 10$ K. Positive bias corresponds to electron injection from the bulk GaAs into the finite superlattice.

wise, S3 shows similarly good agreement.[8] High-voltage deviation may indicate a zero-current model is no longer valid.

The absence of structure in S1 and S2 implies that we have experimentally observed the transition (in this system) from an indistinguishable miniband to a coupled-well structure. In energy, this implies the transition occurs between state splittings of 4 meV (the maximum in S1) and 8 meV (the minimum observable in S3), when $kT <$ the state splitting $E(i+1) - E(i)$. Note that this is a function of the position of eigenstate i within the miniband. In rationalized units, this corresponds to 2.5 meV $< W$(miniband width)$/n$(# periods) $<$ 10.5 meV. The origin of the eigenstate broadening mechanism (such as epitaxial or alloy fluctuations) is not known.

To check that the resonances are indeed arising from the collector density of states, a sample (S5) identical to S4 but with bulk GaAs on one side of the superlattice was investigated. Figure 5 shows the *G-V* characteristics of this structure at 10 K. Positive bias corresponds to electron injection from the bulk GaAs into the finite superlattice. Under this condition, the position and number of the subresonance peaks compares well with that of the finite superlattice injector sample. As has been pointed out earlier,[11] there is no structure in the reverse bias direction since the collector is bulk. It should be noted that the lower Fermi level in the bulk GaAs (versus the replaced superlattice) accounts for the voltage shift of the subresonant peaks.

In summary, we have observed by vertical tunneling transport the eigenstates of a finite superlattice system far below the Stark localization threshold.

We are thankful to R. T. Bate, D. C. Collins, and C. Fonstad for constant support and encouragement, to W. M. Duncan for the photoluminescence measurements, to J. N. Randall for discussions, to H.-L. Tsai for cross-section TEMs, and to R. K. Aldert, P. Q. Montague, E. D. Pijan, P. F. Stickney, F. H. Stovall, and J. R. Thomason for technical assistance. This work was done under the MIT-TI VI-A Internship Program, and was supported in part by the Office of Naval Research.

[1] K. K. Choi, B. F. Levine, R. J. Malik, J. Walker, and C. G. Bethea, Phys. Rev. B **35**, 4172 (1987).
[2] T. Duffield, R. Bhat, M. Koza, F. DeRosa, D. M. Hwang, P. Grabbe, and S. J. Allen, Jr., Phys. Rev, Lett. **56**, 2724 (1986).
[3] B. Deveaud, J. Shah, T. C. Damen, B. Lambert, and A. Regeny,Phys. *Rev. Lett.* **58**, 2582 (1987).
[4] A. Sibille, J. F. Palmier, H. Wang, and F. Mollot, Phys. Rev. Lett. **64**, 52 (1990).
[5] R. A[illegible]vies, M. J. Kelly, and T. M. Kerr, Phys. Rev. Lett. **55**, 1114 (1985).
[6] M. A. Reed, W. R. Frensley, W. M. Duncan, R. J. Matyi, A. C. Seabaugh, and H.-L. Tsai, Appl. Phys. Lett. **54**, 1256 (1989).
[7] J. H. Luscombe, R. J. Aggarwal, M. A. Reed, W. R. Frensley, and M. Luban (unpublished).
[8] The temperature dependence of S2 is nontrivial due to the screening of the superlattice contacts. A full discussion of the transport, characterization, and analysis of structures S1-S4 can be found in R. J. Aggarwal, Master's thesis, MIT, 1990.
[9] R. F. Kazarinov and R. A. Suris, Fiz. Tekh. Poluprov. **5**, 797 (1971) [Sov. Phys. Semicond. **5**, 707 (1971)].
[10] L. L. Chang, L. Esaki, A. Segmüller, and R. Tsu, in *Proceedings of the Twelfth International Conference on the Physics of Semiconductors*, edited by M. H. Pilkuhn (B. G. Teubner, Stuttgart, 1974), p. 688.
[11] P. England, J. R. Hayes, J. P. Harbison, D. M. Hwang, and L. T. Florez, Appl. Phys. Lett. **53**, 391 (1988).